OCEAN ML Model for Transport Risk Mitigation

1. Challenge description
2. Data exploration and preprocessing
    1. Preprocessing of sensor readings data
    2. Vehicles statistics and selection
    3. Features categorization
3. Analytics
    1. Correlations of weather features with sensor targets
    2. Average values of humidity and temperature by the date features
    3. Average values of humidity and temperature by the geolocation groups
    4. The risk of dangerous conditions by the most important features
    5. The risk of dangerous conditions by the specific weather events
    6. Minimizing the risk of dangerous conditions on the long-distance events
    7. The best external weather conditions
4. Modeling
    1. Model for predicting current median temperature inside vehicle
    2. Model for predicting current median humidity inside vehicle
    3. Apply models to the full dataset
    4. Data preprocessing for predicting future target values inside vehicle
    5. Models for predicting 5 future minutes of target values inside vehicle
5. Results and conclusion

1. Challenge description

The objective is to develop ML models using the provided dataset, to predict potentially dangerous temperature and humidity conditions while transporting high-value goods, including livestock and perishable food items. By analyzing telematics, weather data, and sensor readings, the model will uncover the influence of external weather, geolocation, and internal microclimate on transport conditions.

The primary goal is to forecast instances where temperatures may exceed 25°C and/or humidity levels may surpass 80%. Participation in this project offers the opportunity to win prizes of up to $2,500, and there are additional incentives for sharing your work on the Ocean Market. The evaluation of the model will consider factors such as its predictive accuracy, interpretability, and feature selection.
Click here to download data.

Analytics (50 points):

- Find correlations between geolocation, external weather, timestamp, and internal microclimate data. How are these pillars interconnected? (10 points)
- How does this interconnectedness influence the risk of dangerous temperatures or humidity? (10 points)
- How do specific weather events (like high winds) or variables such as the proximity of farms to roads contribute to temperature risk? (10 points)
- How can the risk of temperature or humidity-related issues be minimized for long-distance transportation events? (10 points)
- What combination of internal and external weather factors conclude an ideal temperature that is less than 25 degree celsius and less than 80% humidity? (10 points)

Prediction Model (30 points):
Build a model taking as input data from the "Sensor-Fusion Data" dataset to determine the current indoor temperature and humidity (the "value" column in the "Microclimate Sensor Readings" dataset). Once the model has been created, use it to simulate a trip of your choice. You will be judged on your choice of features and model.

Optional Bonus: Prediction Model (20 points)
Build a model that predicts indoor temperature and humidity a few minutes/seconds into the future. This requires data transformation. If you can create a model, show us how!

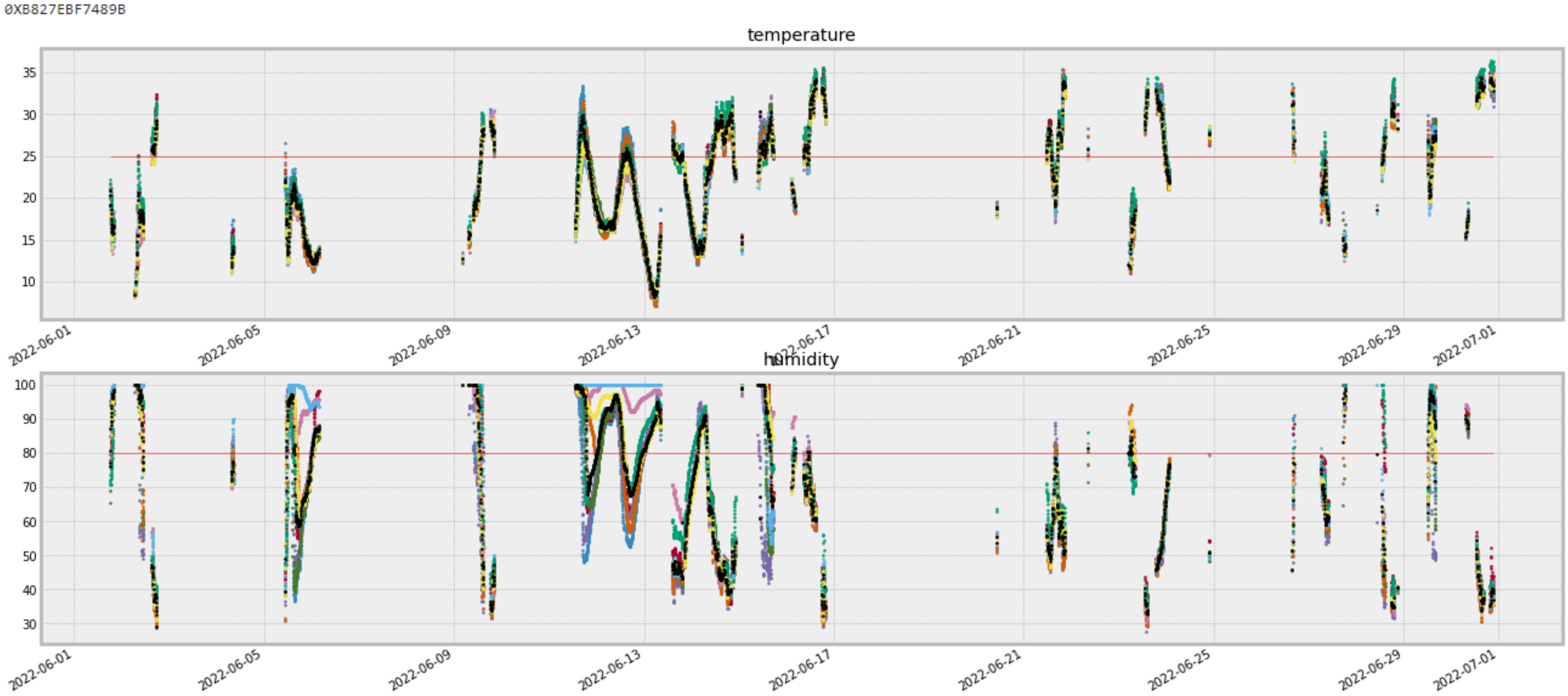
Report (20 points):
Submit a report describing the above findings. Make sure to include qualitative insights in addition to quantitative ones. Reports will be evaluated on presentation structure, approach, content, and completeness.

2.1. Preprocessing of sensor readings data

On the picture below shown internal microclimate data – sensors measurements of temperature and humidity selected for one vehicle in June 2022. We observe that at the same point of time the temperature and humidity values measured by different sensors are different (they marked by different colors). To aggregate that values into one value we calculated the medians of temperature and humidity by each unique timestamp and used them as the main target variables. On the picture shown the 5-minute median values as black dots.

By the red lines shown the cutoffs above which transportation conditions becomes dangerous. We created categorical variables to indicate these cases:

- *target_humidity_cat* – 1 if median humidity > 80% **(72% of cases)**;
- *target_temperature_cat* – 1 if median temperature > 25°C **(16% of cases)**;
- *target_cat* – 1 if *target_humidity_cat* = 1 OR *target_temperature_cat* = 1 **(84% of cases)**;

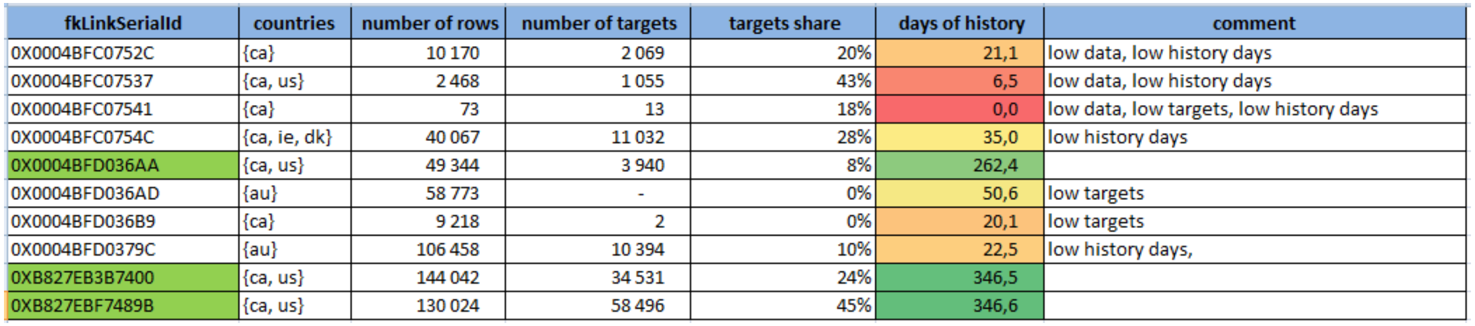


* More details you can find in the script: https://github.com/abessalov/Ocean_Transportation/blob/master/1_preprocessing.ipynb

2.2. Vehicles statistics and selection

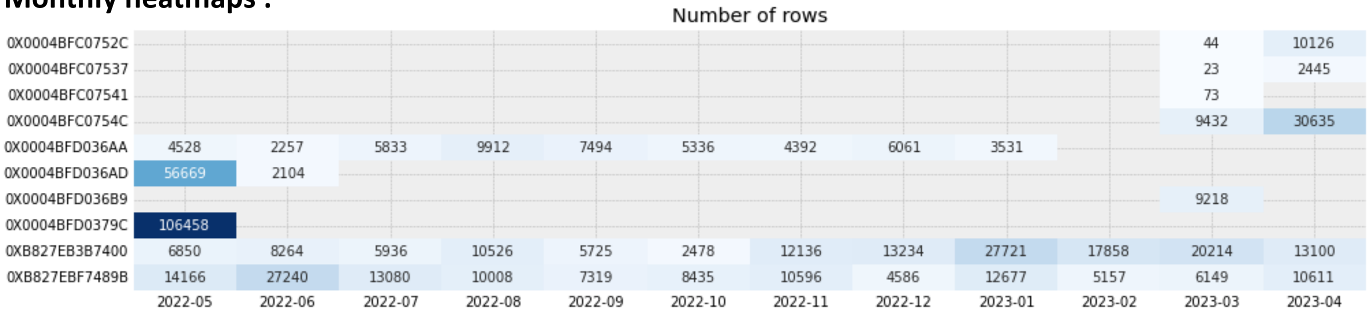
All our input dataset consists of 10 unique vehicles. From the table and heatmaps below we decided to choose 3 vehicles travelled in Canada and US for whom we have enough data for further analytics and modeling (marked as green in the table). They have a data practically for the full year so we can identify seasonal dependencies, find correlations between weather variables and target measurements.

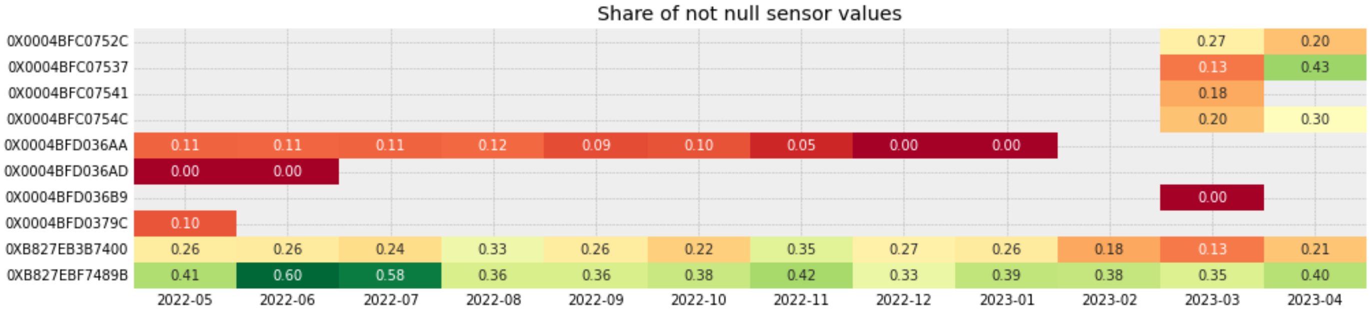
Vehicles statistics:



fkLinkSerialId	countries	number of rows	number of targets	targets share	days of history	comment
0X0004BFC0752C	{ca}	10 170	2 069	20%	21,1	low data, low history days
0X0004BFC07537	{ca, us}	2 468	1 055	43%	6,5	low data, low history days
0X0004BFC07541	{ca}	73	13	18%	0,0	low data, low targets, low history days
0X0004BFC0754C	{ca, ie, dk}	40 067	11 032	28%	35,0	low history days
0X0004BFD036AA	{ca, us}	49 344	3 940	8%	262,4	
0X0004BFD036AD	{au}	58 773	-	0%	50,6	low targets
0X0004BFD036B9	{ca}	9 218	2	0%	20,1	low targets
0X0004BFD0379C	{au}	106 458	10 394	10%	22,5	low history days,
0XB827EB3B7400	{ca, us}	144 042	34 531	24%	346,5	
0XB827EBF7489B	{ca, us}	130 024	58 496	45%	346,6	

Monthly heatmaps :

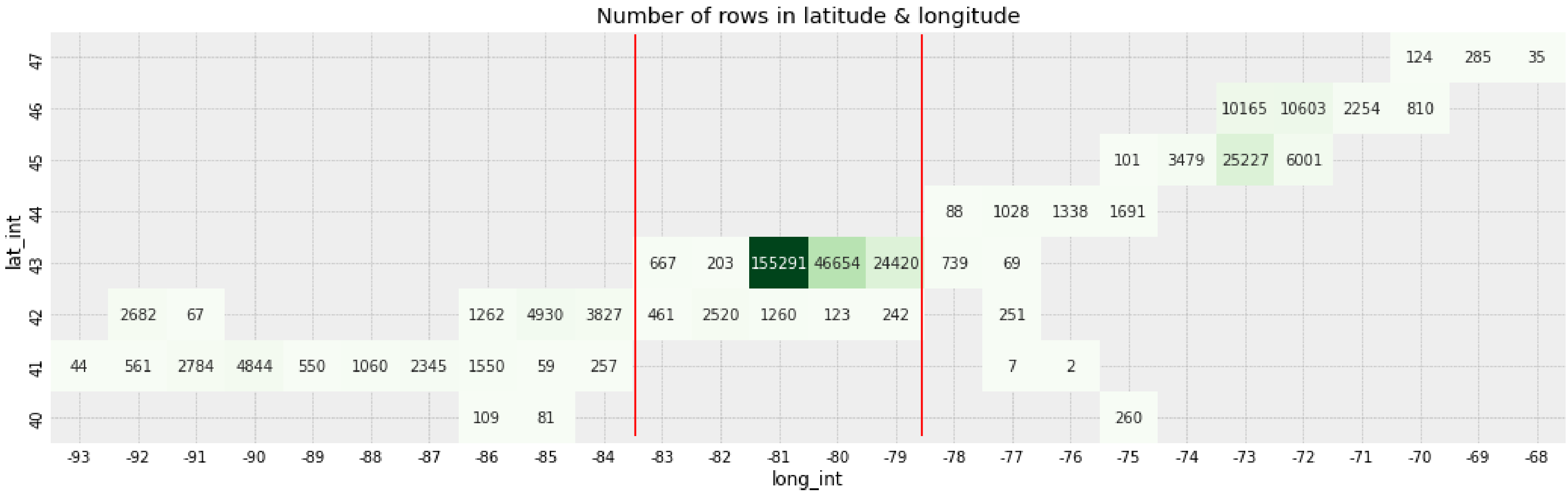




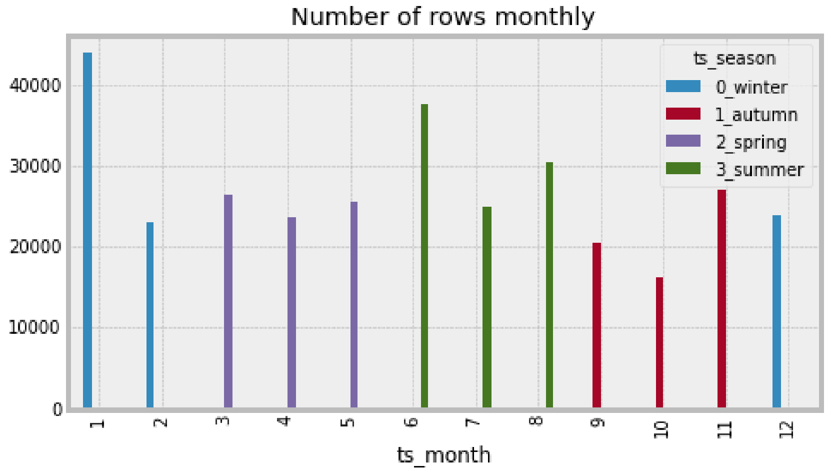
* More details you can find in the script: https://github.com/abessalov/Ocean_Transportation/blob/master/1_preprocessing.ipynb

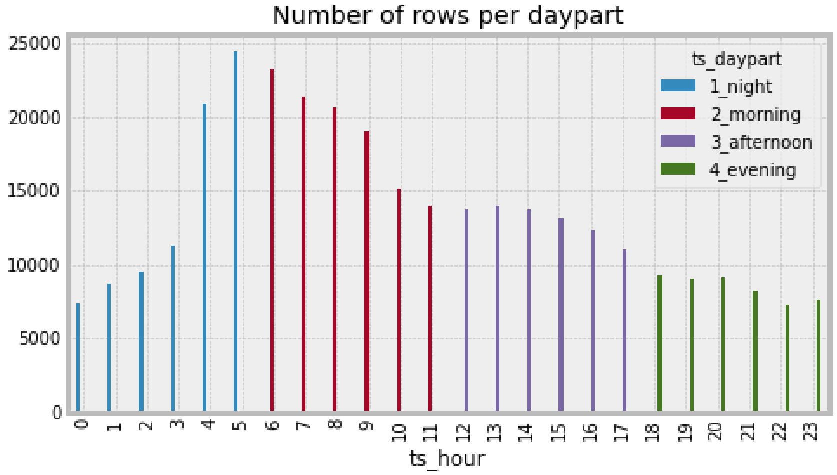
2.3. Features categorization

Let's look at the geographical coordinates of our selected vehicles. On the picture below we shown the number of rows depending on the latitude and longitude. We can see that all points lie practically on one line so for further analytics let's create the grouping variable based on the longitude only (the splits marked as red lines):


Number of rows in latitude & longitude

Also it will be useful to create seasonal variables: year season and part of the day:


Number of rows monthly

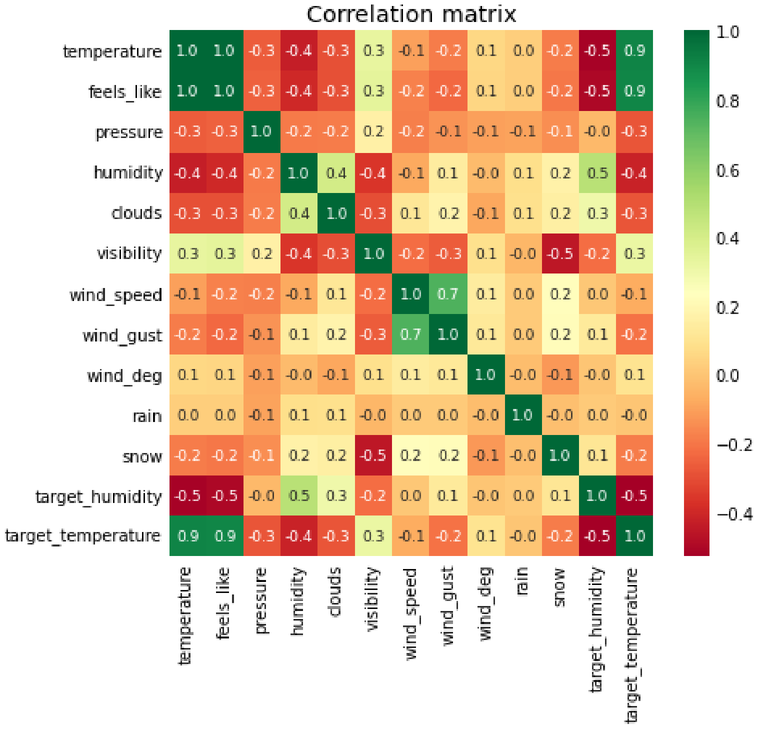

Number of rows per daypart

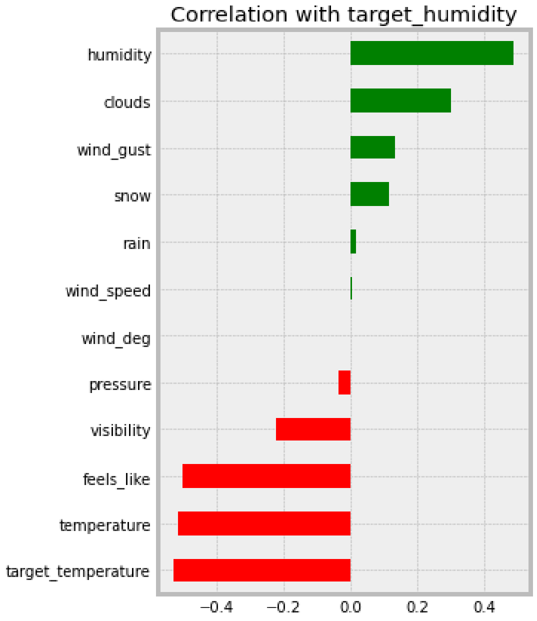
* More details you can find in the script: https://github.com/abessalov/Ocean_Transportation/blob/master/1_preprocessing.ipynb

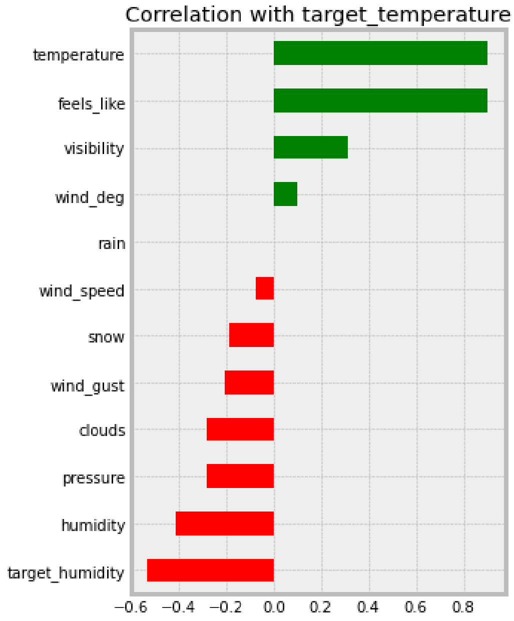
3.1. Correlations of weather features with sensor targets

On the pictures below we shown the Pearson correlation coefficient of weather features with microclimate sensors targets – temperature and humidity. We can get some facts from here:

- There is a high positive correlation between temperature and target_temperature (median of sensor temperature readings) ~0.9.
- There is a medium positive correlation between humidity and target_ humidity (median of sensor humidity readings) ~0.5.
- Humidity and temperature have negative medium correlation ~-0.4.
- Some weather features have high correlation between each other, for example: temperature and feels like, wind speed and wind gust.







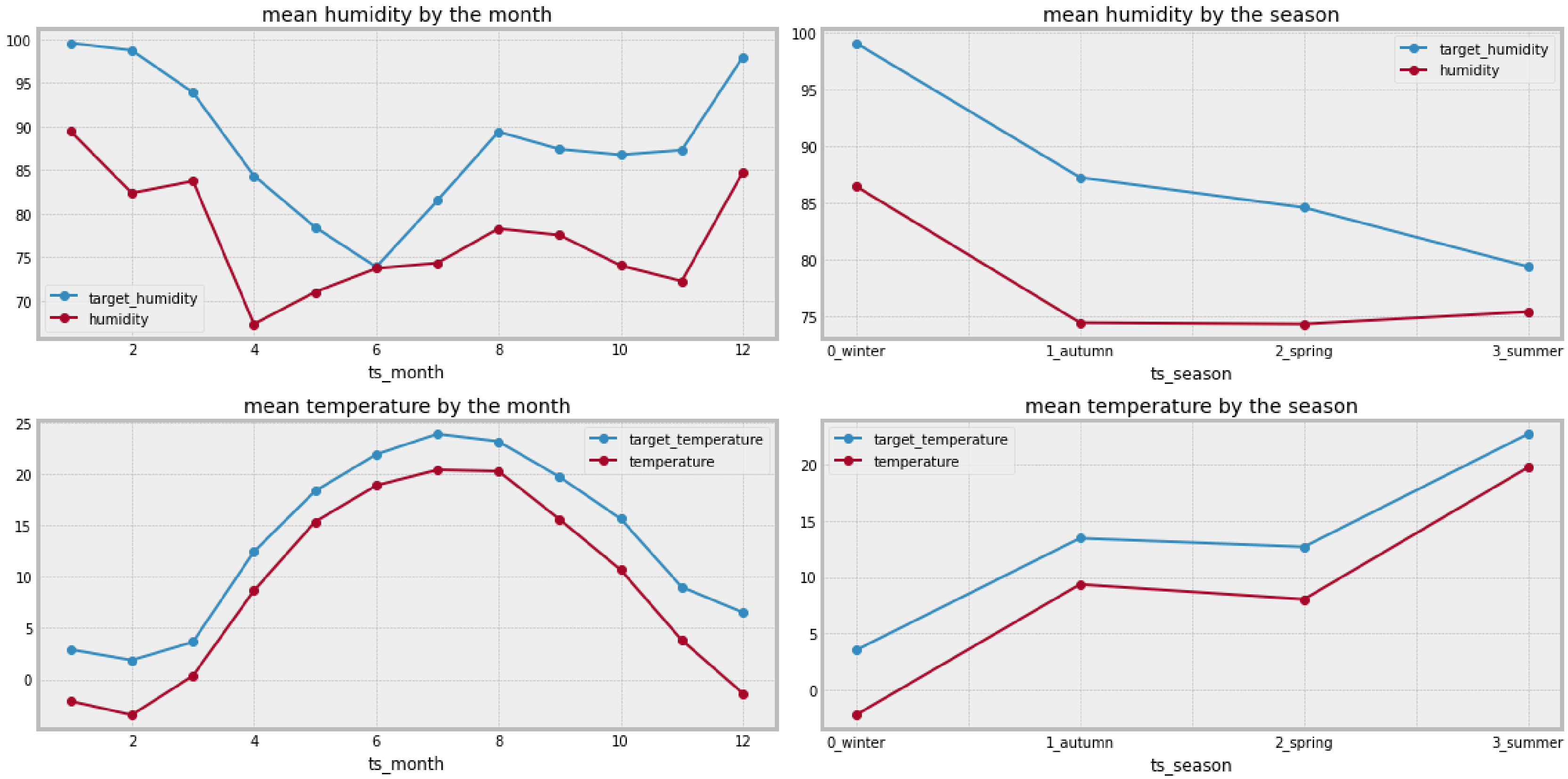
* More details you can find in the script: https://github.com/abessalov/Ocean_Transportation/blob/master/2_statistics.ipynb

3.2. Average values of humidity and temperature by the date features

We established that outside temperature have good positive correlation with inside value and the same is true for the humidity. Now let's compare their average values by the date features – month and season to identify seasonal patterns:

- The highest values of the temperature are in the summer time, the lowest values – in the winter.
- On the bottom pictures we can see the strength of the correlation between inside and outside temperatures.
- The highest value of humidity in the winter time, the lowest inside humidity in the summer. Average outside humidity is practically the same for all seasons except winter.

Average values by the months and seasons:



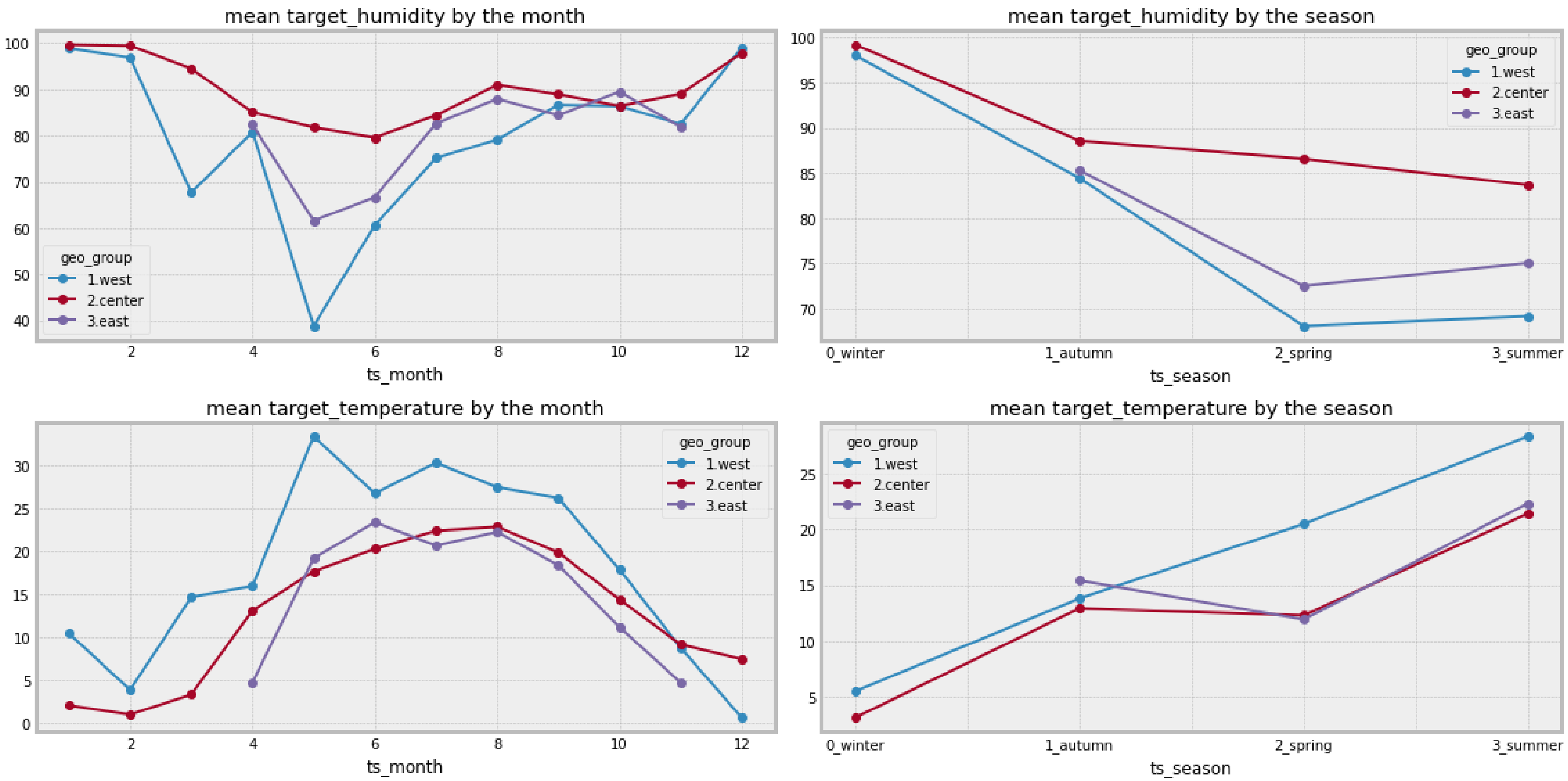
* More details you can find in the script: https://github.com/abessalov/Ocean_Transportation/blob/master/2_statistics.ipynb

3.3. Average values of humidity and temperature by the geolocation groups

We shown the similar picture as before but with splitting by the geolocation group: west, center, east:

- The highest average values of the inside temperature are on the west geolocation.
- Average values of inside temperature on the west and center are practically the same.
- The highest average values of the inside humidity are in the central geolocation.

Average values of target values in different regions:



mean target_humidity by the month

mean target_humidity by the season

mean target_temperature by the month

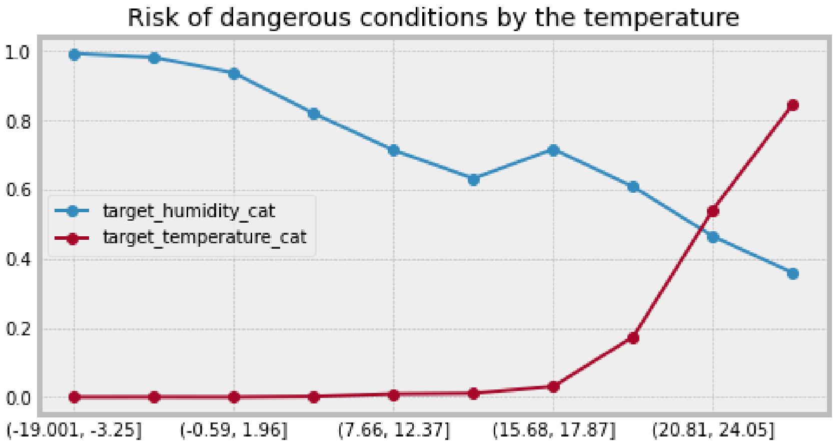
mean target_temperature by the season

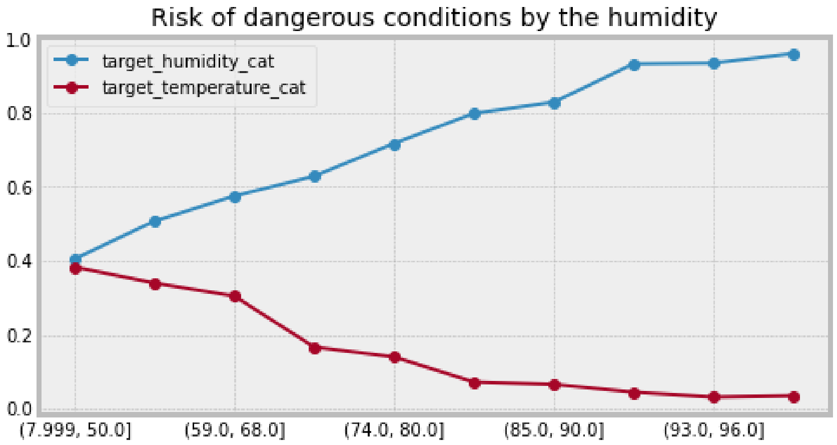
* More details you can find in the script: https://github.com/abessalov/Ocean_Transportation/blob/master/2_statistics.ipynb

3.4. The risk of dangerous conditions by the most important features

For further analysis we will use the risk of dangerous conditions that we defined as the number of cases when the corresponding feature (inside temperature, humidity or any) exceeds the norm (25°C for temperature, 80% for humidity).
On the picture below shown these values of risks depending on the bucket of the most important outside weather features – temperature and humidity. We can get the simple rules best on that plots:

- If the outside temperature is less then 18°C then practically there is no risk of dangerous temperature inside the vehicle.
- After 18°C there is a rule: the higher the outside temperature the higher the risk of dangerous temperature inside the vehicle.
- The higher the outside humidity, the higher the risk of dangerous humidity inside the vehicle.
- The higher the outside temperature, the lower the risk of dangerous humidity inside the vehicle.
- The higher the outside humidity, the lower the risk of dangerous temperature inside the vehicle.



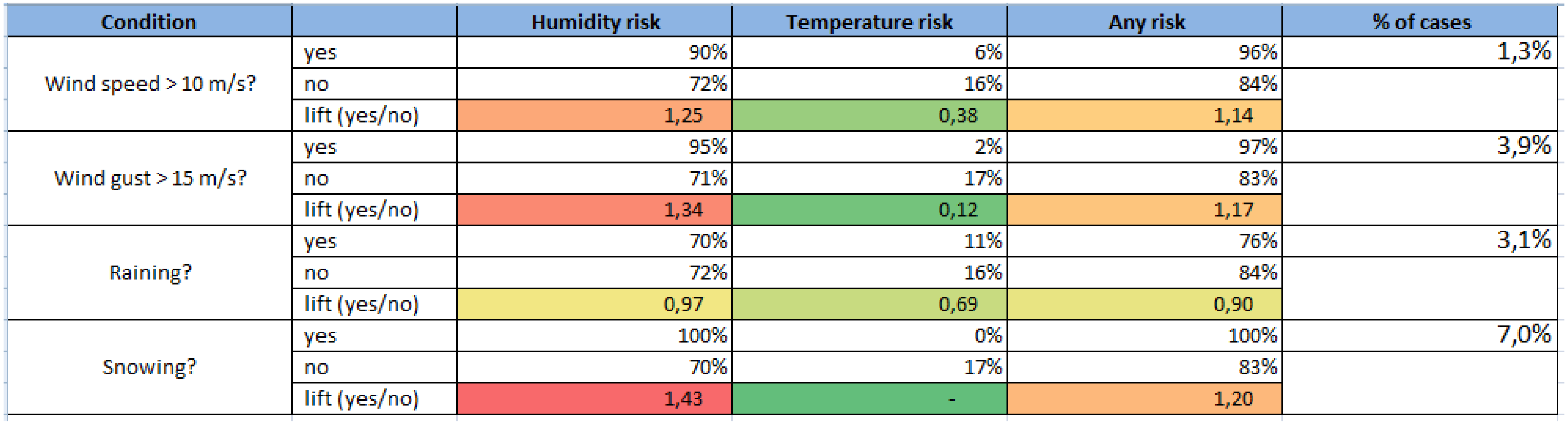


* More details you can find in the script: https://github.com/abessalov/Ocean_Transportation/blob/master/2_statistics.ipynb

3.5. The risk of dangerous conditions by the specific weather events

In the table below shown the probability of the dangerous conditions depending on the rarely specific weather events. Let's give some explanations and build the rules from that table:

- If wind speed is more than 10 m/s (it's about 1,3% of cases) then:
  - the probability of temperature risk is lower (lift < 1);
  - the probability of humidity risk is higher (lift > 1);
- If wind gust is more than 15 m/s (it's about 3,9% of cases) then:
  - the probability of temperature risk is lower (lift < 1);
  - the probability of humidity risk is higher (lift > 1);
- If it's raining (it's about 3,1% of cases) then the probability of temperature risk is lower (lift < 1);
- If it's snowing (it's about 7% of cases) then:
  - there is no temperature risk;
  - the probability of humidity risk is higher (lift > 1)



Condition		Humidity risk	Temperature risk	Any risk	% of cases
Wind speed > 10 m/s?	yes	90%	6%	96%	1,3%
	no	72%	16%	84%	
	lift (yes/no)	1,25	0,38	1,14	
Wind gust > 15 m/s?	yes	95%	2%	97%	3,9%
	no	71%	17%	83%	
	lift (yes/no)	1,34	0,12	1,17	
Raining?	yes	70%	11%	76%	3,1%
	no	72%	16%	84%	
	lift (yes/no)	0,97	0,69	0,90	
Snowing?	yes	100%	0%	100%	7,0%
	no	70%	17%	83%	
	lift (yes/no)	1,43	-	1,20	

* More details you can find in the script: https://github.com/abessalov/Ocean_Transportation/blob/master/2_statistics.ipynb

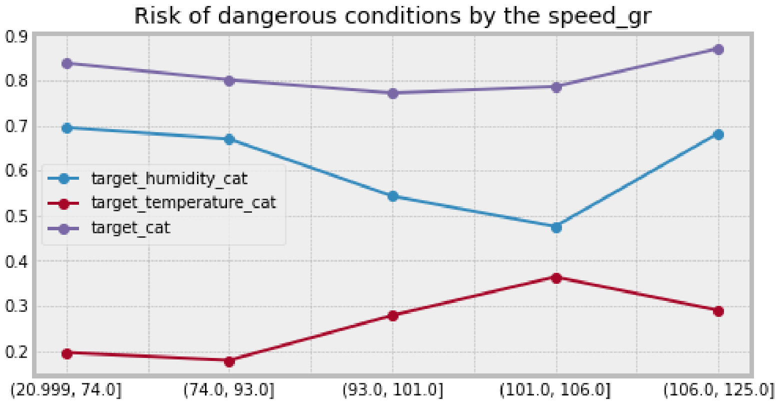
3.6. Minimizing the risk of dangerous conditions on the long-distance events

We selected the long-distance events as the rows where speed > 20 km/h and have done the following analytics based on that events.

What is the optimal speed to minimize risk:

Let’s create 5 bins of speed feature and calculate the share of rows with the dangerous conditions (red line – for temperature, blue line – for humidity, purple line – for all). From the right picture:

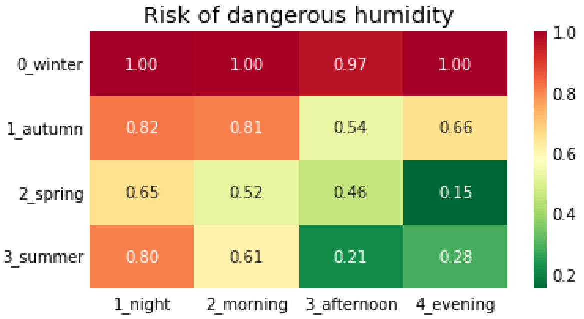
- The optimal speed for minimizing temperature is below 93 km/h,
- The optimal speed for minimizing humidity is 93-106 km/h.

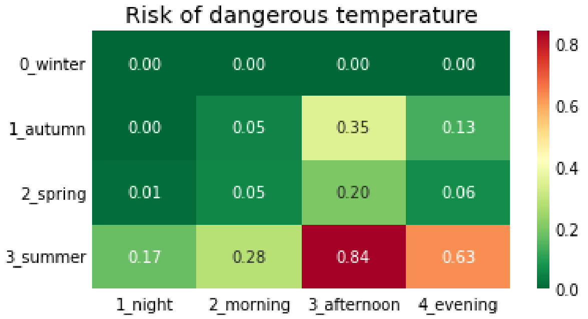


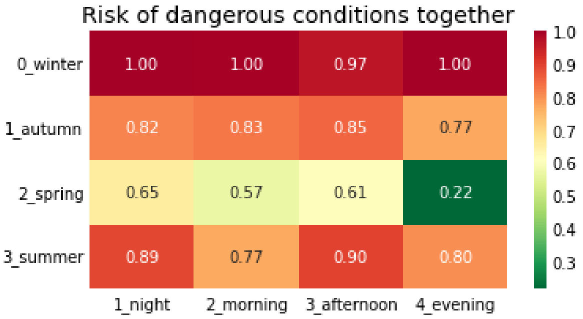
What is the optimal time for transportation depending on season:

We can highlight the interesting facts from the picture below:

- For winter season there is no risks of dangerous temperature, humidity practically everywhere is equal 1, that’s strange.
- For other seasons maximum risk of dangerous temperature is in the afternoon.
- For other seasons maximum risk of dangerous humidity is in the night time.
- On the last right heatmap shown the optimal times minimizing the risk of any dangerous situation:
    - The best transportation time in the autumn and spring seasons is evening,
    - The best transportation time in the summer is morning.







* More details you can find in the script: https://github.com/abessalov/Ocean_Transportation/blob/master/2_statistics.ipynb

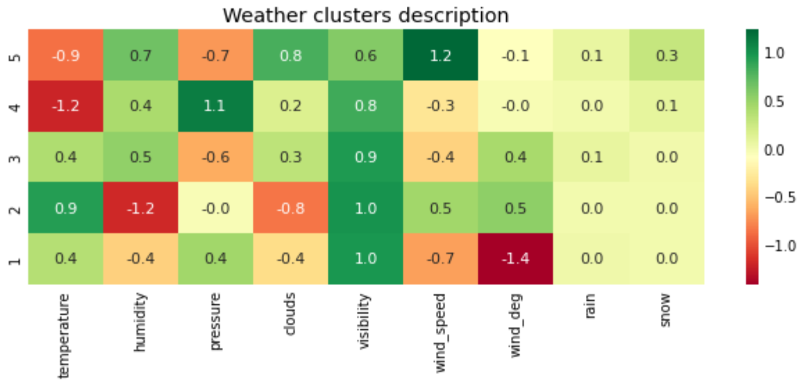
3.7. The best external weather conditions

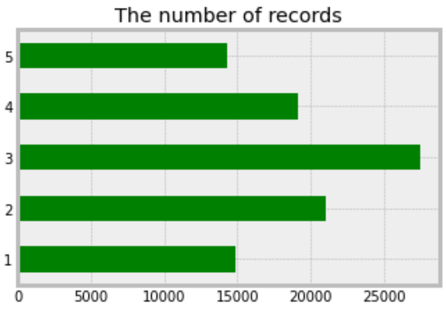
Now we want to understand what combination of weather factors conclude an ideal internal temperature and humidity for transportation. We've made this analytics by the following steps:

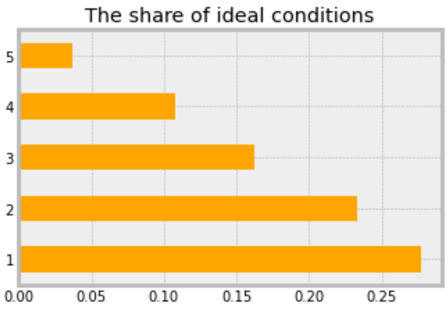
- Explore the weather features distributions and correlation matrix and do some preprocessing:
  - remove highly correlated features – feels like and wind gust,
  - remove outliers by clipping the values to some ranges,
  - features binning,
  - normalization by transforming to z-scores.

- Apply K-means clustering model with K = 5 and interpret clusters received. On the picture below you can see the average z-scores values in every cluster – we can highlight some important characteristics of each cluster:
  - cluster 5 – low temperature, high humidity, high wind, cloudy, probability of snow, rain;
  - cluster 4 – low temperature, high humidity, low wind, probability of snow;
  - cluster 3 – medium temperature, high humidity, low wind, probability of rain;
  - cluster 2 – high temperature, low humidity, medium wind, no snow, no rain;
  - **cluster 1 – medium temperature, medium humidity, low wind, no snow, no rain;**

- Generate a feature indicating the ideal internal weather conditions and calculate the share of these cases in each cluster to identify the best ones. You can see these shares on the right picture below, it means that cluster 1 is the best for transportation.







* More details you can find in the script: https://github.com/abessalov/Ocean_Transportation/blob/master/2_statistics.ipynb

4.1. Model for predicting current median temperature inside vehicle

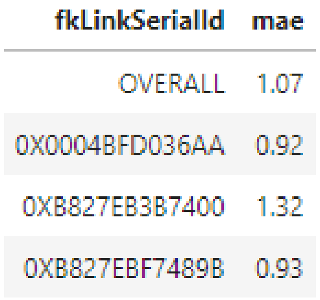
Task description:
- There are a lot of missing values in the temperature measurements from the *left dataset* that we need to fill.
- Other features from that dataset have values therefore we have chosen it as the main for 2 tasks:
  - Develop a model for predicting medians of inside temperature at each point of time.
  - Apply this model to the rows where the temperature measurements unknown.

Validation set: 25% randomly sampled rows.
Training set: The rest 75% of rows.

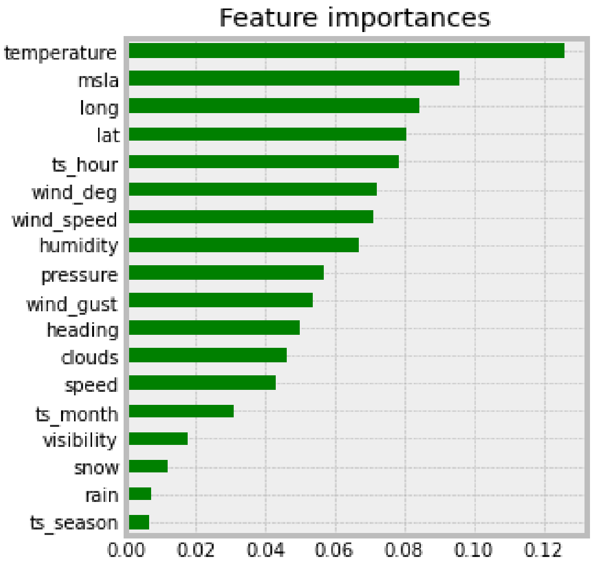
Target: Median of sensor temperature measurements at the same point of time.
Features:
- 10 weather features (temperature, pressure, humidity, clouds, visibility, wind_speed, wind_gust, wind_deg, rain, snow)
- 5 geolocation features (lat, long, msla, heading, speed)
- season, month and hour of the transaction

Model : Gradient Boosting (LightGBM)
Quality metric: MAE (mean absolute error)

Results:



fkLinkSerialId	mae
OVERALL	1.07
0X0004BFD036AA	0.92
0XB827EB3B7400	1.32
0XB827EBF7489B	0.93



Comments:
- We received very good overall MAE value =1.07 degree celcıus.
- Two vehicles have practically the same results, one vehicle worse on about 0.4.
- We also tried to train individual models for each vehicle but the results are practically the same hence we don’t need to have separate models for each vehicle and one universal model is enough.
- Temperature is the most important feature for the model, geolocation features are also very important.

* More details you can find in the script: https://github.com/abessalov/Ocean_Transportation/blob/master/3_modelling_regression.ipynb

4.2. Model for predicting current median humidity inside vehicle

Task description:

- There are a lot of missing values in the humidity measurements from the *left dataset* that we need to fill.
- Other features from that dataset have values therefore we have chosen it as the main for 2 tasks:
  - Develop a model for predicting medians of inside humidity at each point of time.
  - Apply this model to the rows where the humidity measurements unknown.

Validation set: 25% randomly sampled rows.

Training set: The rest 75% of rows.

Target: Median of sensor humidity measurements at the same point of time.

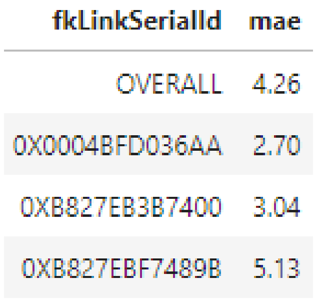
Features:

- 10 weather features (temperature, pressure, humidity, clouds, visibility, wind_speed, wind_gust, wind_deg, rain, snow)
- 5 geolocation features (lat, long, msla, heading, speed)
- season, month and hour of the transaction

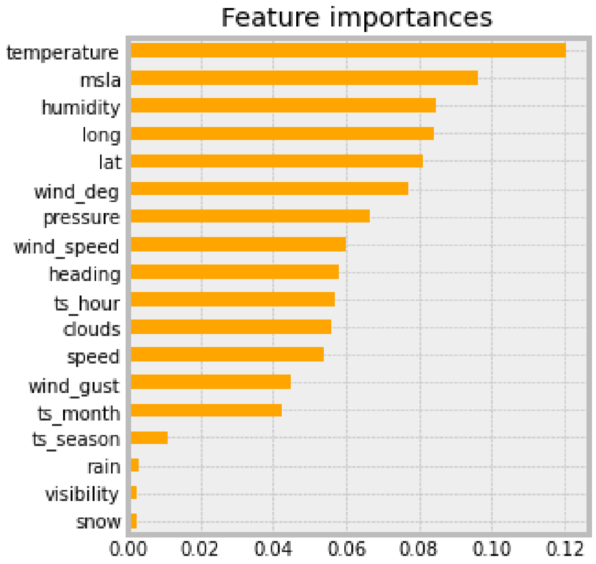
Model : Gradient Boosting (LightGBM)

Quality metric: MAE (mean absolute error)

Results:



fkLinkSerialId	mae
OVERALL	4.26
0X0004BFD036AA	2.70
0XB827EB3B7400	3.04
0XB827EBF7489B	5.13



Comments:

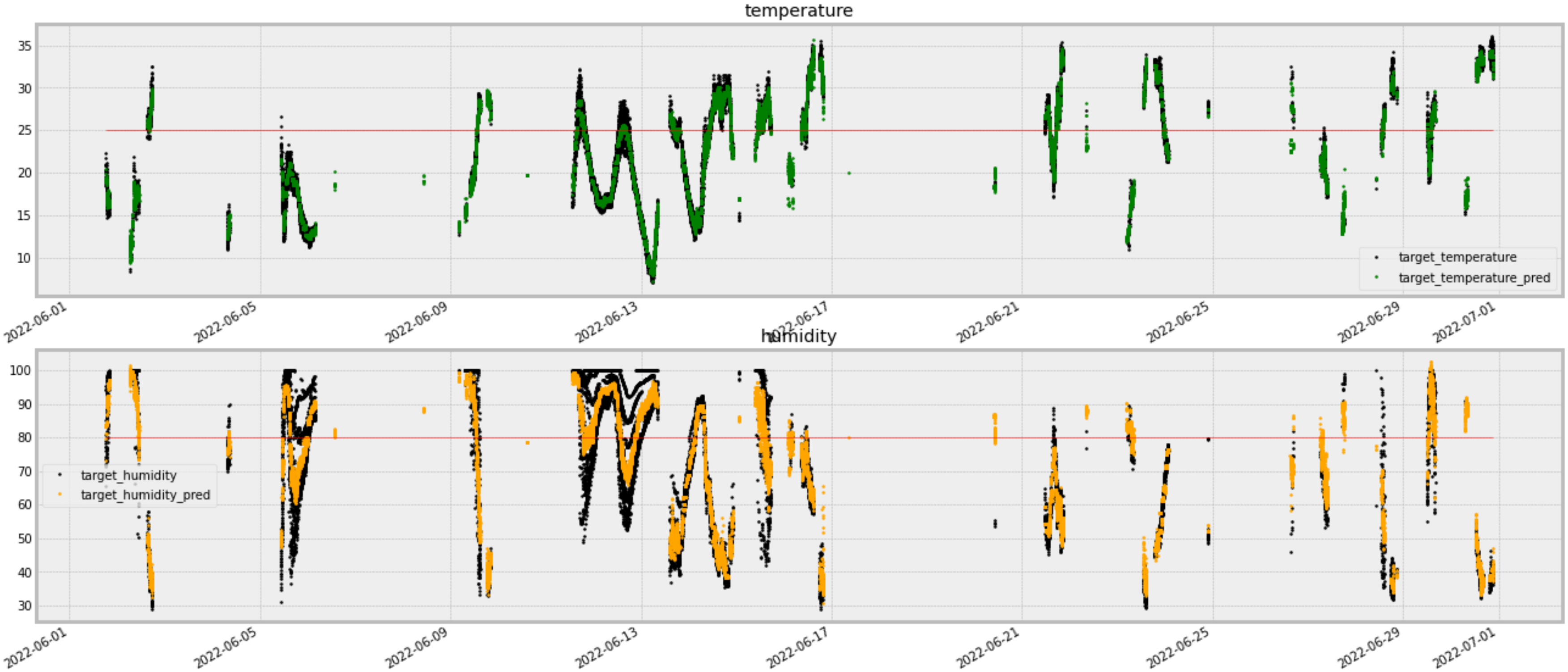
- We received rather good overall MAE value = 4.26% of humidity.
- One vehicle have significantly worse result than others.
- We also tried to train individual models for each vehicle but the results are practically the same hence we don't need to have separate models for each vehicle and one universal model is enough.
- Like in the previous model temperature is the most important feature and geolocation features are also very important.

* More details you can find in the script: https://github.com/abessalov/Ocean_Transportation/blob/master/3_modelling_regression.ipynb

4.3. Apply models to the full dataset

We applied developed models to fill missing values of target measurements of inside temperature and humidity. The figure below show the results for one vehicle in June 2022 (as we showed earlier without predictions). We can highlight the interesting facts from here:

- Real values and predictions looks very similar that is proving the good quality of models built.
- Forecasts looks smoother. For example, by looking at the humidity the real points jumps a lot and it seems that there can be several unique values for the same point of time, but it's not, these are actually values at different times.
- The model smoothes such situations and give good predictions which we will use for further analysis.

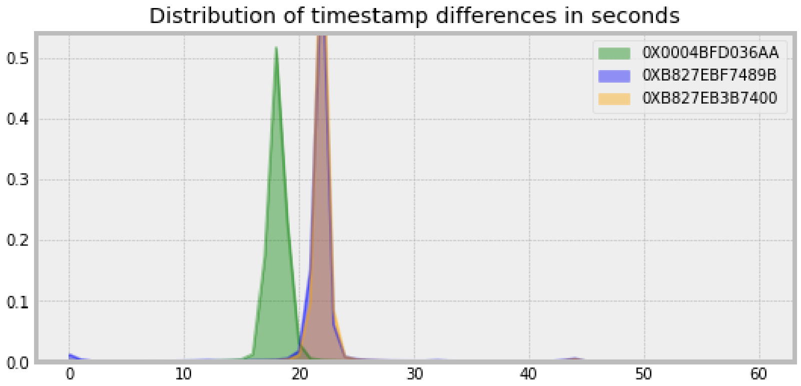


* More details you can find in the script: https://github.com/abessalov/Ocean_Transportation/blob/master/3_modelling_regression.ipynb

4.4. Data preprocessing for predicting future target values inside vehicle

Task description:

- By looking at the distribution of time difference between measurements on the picture below we observe that there are peak values in 17-19 seconds for the 0X0004BFD036AA vehicle and in 21-23 seconds for the rest.
- Also about 95% of data records lies in the 1-minute period between each other, the rest may indicate the turning off sensors, vehicles stopping etc.
- To do some universal interval between measurements we will resample our data to 1-minute interval and calculate median values of the records for each minute.
- And now let's formulate the task the following way: based on the previous values of temperature and humidity we need to predict their every minute values for 5 minutes ahead.



Data preprocessing:

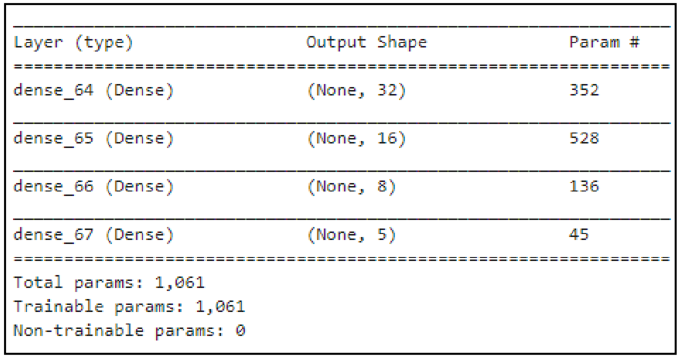
- We took only one vehicle 0XB827EBF7489B and made a median resampling of predicted temperature and humidity values into 1 minute interval.
- We have generated a sequences of 15 minutes measurements without interruption and broke them into 10 for input, 5 for output.
- The data rows have been normalized by calculating z-scores.
- As a result, we got input and output tensors of dimensions (39130, 10) and (39130, 5) for models developing.

* More details you can find in the script: https://github.com/abessalov/Ocean_Transportation/blob/master/4_modelling_ts.ipynb

4.5. Models for predicting 5 future minutes of target values inside vehicle

Model description:

- We have built individual models for predicting future values of temperature and humidity based on their previous ones.
- The last 25% of 5 minutes intervals were used as validation set, the rest for training.
- We have tested neural networks consisting of Dense and Dropout layers and took the best model in terms of MAE metric on the validation sets and also considering model complexity. On the picture below shown the best architecture both for the temperature and humidity.
- In the training process we were using 200 number of epochs with batch size = 1024. Optimizer is RMSprop(lr=0.001) with MAE evaluation metric.

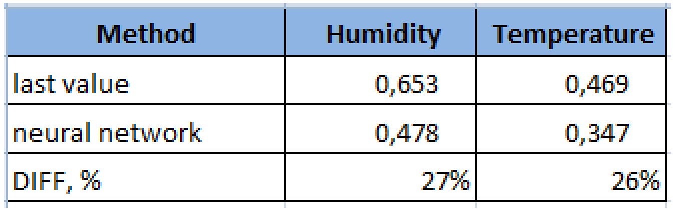


```
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
dense_64 (Dense)             (None, 32)                352       
_________________________________________________________________
dense_65 (Dense)             (None, 16)                528       
_________________________________________________________________
dense_66 (Dense)             (None, 8)                 136       
_________________________________________________________________
dense_67 (Dense)             (None, 5)                 45        
=================================================================
Total params: 1,061
Trainable params: 1,061
Non-trainable params: 0
```

Results:

On the table below shown the MAE scores on the evaluation set:

- We received a very good overall MAE scores both for temperature and humidity. For temperature MAE ~0.347, for humidity ~0.478.
- By comparing neural network with the predictions made by the last known value we received a sufficiently high increase in quality on 26-27%, that is proving the fact that neural networks is good approach for this task.



Method	Humidity	Temperature
last value	0,653	0,469
neural network	0,478	0,347
DIFF, %	27%	26%

* More details you can find in the script: https://github.com/abessalov/Ocean_Transportation/blob/master/4_modelling_ts.ipynb

5. Results and conclusion

We have analyzed the dataset for transport risk mitigation problem and after exploration can highlight the following interesting facts:

1. We calculated correlations between input features and sensor target measurements of humidity and temperature and can conclude that there is a high positive correlation between outside temperature and inside ~0.9. Correlation between inside and outside humidity is lower and equals ~0.5.
2. We calculated the risk of dangerous conditions inside the vehicle by the different features and found the simple rules and some interesting insights also.
3. We found the optimal time of the day depending on the season time to minimize the risk of dangerous conditions on the long-distance events.
4. We built a clusters by using weather features and found the best factors concluding the ideal conditions inside the vehicle. They are: medium temperature, medium humidity, low wind, no snow, no rain.
5. We developed a gradient boosting models for predicting current vehicle's inside temperature and humidity based on the input vehicle features. We received sufficiently good MAE scores on the validation sets: for temperature ~1.07 °C and for humidity ~4.26% of humidity. These models we used to fill empty values of temperature and humidity when we have the input features but target measurements is unknown.
6. We developed a neural networks models for predicting 5 future minutes of inside temperature and humidity values and received a very good MAE scores: for temperature ~0.347, for humidity ~0.478. By comparing these results with the last known value prediction we got increase in quality on 26-27%. One possible improvement here is try to use more sophisticated neural network architectures like RNN (GRU, LSTM) that we didn't consider in that work.