Aalto Pro, Diploma in Artificial Intelligence

Project work:

Applying data science and machine learning in development of a tool for a waste management system onboard cruise ship

Dr. Jari Jokela, Head of Research
Evac Oy

Abstract

Modern cruise ships are operated by owners that have adapted strict environmental policies in their operations. The time spent in ports of call are getting longer. In addition, the cruise ship customers are taking cruise to a route that have destinations with a significant natural conservation value. This is reflecting the increase of environmental awareness. Thereby, any signs of bad environmental management of a cruise ship will be a significant image risk to the shipowner. One of the biggest sources of pollution load of the cruise ships is the food waste system, that contributes c. 70 % of the organic and nutrient load of a cruise ship.

The purpose of the work is to prototype data science operations with a shipboard process data uploaded data of food waste from a cruise ship. This data was converted and filtered by data science methods to a form that could be used for machine learning methods. In this study we converted the data of two individual food waste systems into production rate over time as Pandas data frames in Jupyter Notebook. These data frames were split into the time spend in the port of call and into the time period that ship was in sea conditions. These datasets were used to predict the future pattern of food waste production in order to help ship manage their waste streams pro-actively to minimize the risk of waste spills.

The production data as measurement of tank level turned out to be extremely noisy data. The reason was the sampling frequency of the ship IoT-system that was 5 seconds. Not only the did the data collection produce huge data files, consisting several files with half million lines but also it emphasized the noises in the data like rolling and pitching of the ship. This caused an excessive filtering. It is utmost important to establish a proper data management system to have smooth running end-to-end data utilization for the business benefits. The data can be after that used for the operations of machine learning.

A predictive model was produced with a polynomial regression method of Scikit-Learn module. The labeled production rate was modelled against the moments of time whilst being in port or at sea. The polynomial model gave fairly good fitting to the data. The error of the model was computed against the training data. The model validation was done with a r2_score method of the Scikit-learn-package. The model did not perform well. More prototyping of the model development needs to be done in order to produce reliable tool for prediction.

Introduction

Background

Cruise ship waste

A cruise ship is a passenger ship used for pleasure voyages when the voyage itself, the ship's amenities, and sometimes the different destinations along the way (i.e., ports of call), form part of the passengers' experience [1]. Modern cruise ships can be considered as a combination of hotel, shopping center, amusement park, casino, spa and a center of all kind of restaurants with big galleys for preparation of all kind of meals. Inclusion and development of these activities is a service evolution of several decades that is basically a desire of each cruise ship owner to create bigger and better hence more attractive series of new cruise ships for potential visitors.

All these onboard activities produce high volumes of wastes that is around 2 to 3 times the amount when compared to an average amount produced by municipal centers per capita in the Western countries [2]. Especially, as all the meals (around 8 per each day) are prepared onboard the ship in the central galleys. In addition to central galleys, there are more than ten small kitchens, pantries and bars around the ship. Also, there are dedicated messes for ship crew. The amount of crew adds c. 1/3 manning to the amount of ship customers. Thereby, organic and nutrient loading of the waste, if discharged to the sea, is very high.

The waste management system of a modern cruise ship is confined to ship machinery spaces. There are basically three type of solid waste produced onboard: Dry waste (DW), food waste (FW) and biosludge waste (BW). The DW is mainly solid packing materials from the provisions and unsorted waste from the cabins and restaurants. The DW contains mainly plastics, cardboard, paper and packing materials. The FW is produced mainly from the food preparation from the gallies, but also as a food rejects from bars, restaurants and crew messes of the ship. Food waste is the main source of organic loading and nutrients from the ship. The BW is a biological surplus sludge from the ship wastewater treatment plant.

The management and the discharge of the waste materials is strictly regulated by both international and national legislations and guidelines. International Maritime Organisation (IMO) [3] The discharge of garbage is strictly prohibited in all sea areas. The discharge of FW and BW is also strictly

regulated and also prohibited in many special sea areas. In addition to this the storage silos for these waste materials is very limited and is sufficient only for couple of days. Until the recent days, disposal of the wet waste (FW and BW) has been a common practice. The main limitation (by IMO) has been prohibiting the discharge on the shore area (within 12 nm) and in so-called special sea areas. Nowadays, owing to the pressure from the customers, the direct disposal to sea is no longer practiced. However, it is not prohibited in the sea conditions.

Motivation of the work

Evac Oy supplies integrated cleantech solutions, including e.g. waste management systems, to all types of ships. The cruise ships are our biggest market segment and for past few years Evac has had more than 80 % of market share in waste systems. The sustainability is the key value for all major cruise ship owners. Not only because of regulations and guidelines, but because of acceptability of the cruise business as a whole. Basically, the cruise ships are operated by zero-discharge principle.

The cruise ships are operated on a strictly defined itinerary and in the given schedule. The shipboard waste management is not the main operation of the ship and should, by no means, affect the operation of a ship. On the other hand, volumetric production of the waste streams (DW, FW, BW) are not constant and are fairly unpredictable. As mentioned, the onboard storage silos are very small in volume. For that reason, cruise ships are many times forced to store waste container in corridors and luggage storages. For the same reason, ships are frequently discharging high proportion of untreated FW and BW directly into the sea. This poses high risk for the reputation of the ship owners as well as for the whole cruise business.

Purpose and objective

Evac has no experience in data collection and in utilization of data. Over 30 000 reference ships are all disconnected. Evac started an ambitious digitalization project year ago in order to produce added value from the onboard systems from our customer ships. The purpose of this work is to do first piloting of data utilization from one of Evac customer ship. This will be done by using data science tools. Baring in mind that our company has no background in working with data, this is significant step forward into digital business.

In this work, datasets from shipboard IoT-system is used for data science prototyping. The primary purpose is to demonstrate usefulness of data science tools for marine environmental protection. This work will produce valuable information how to develop data infrastructure and data management and data science tools further in the future. Objective is to study and prototype use of raw data to see what kind of programming is needed to get the data into useful form. In this work, an exploratory analysis on the waste data produced from one of the ships of TUI Cruises to produce an insight of the dynamics of the waste production. Finally, this cleaned and organized data in form of dataframe is used for

prototyping with machine learning tools. This would be a simple prediction task by using e.g. linear regression modules of Scikit-learn module.

The primary goal is to make first data science piloting in a company with no background in data usage. This first piloting will provide initial methods how to extract usable data from the existing sources, clean and explore the data so that a usable ML models could be produced and trained with new and fresh data in the future from the IoT-system. It is important to note that an end-to-end data science and machine learning production pipeline is not realistic. However, this work will provide a foundation for that goal by providing guidelines how to carry out data science methods. This in practice contain:

- What kind of data science programming environment to use
- What kind of data management actions are needed
- Produce the first library of code snippets to clean, filter and explore the shipboard data
- Use of data with a predictive machine learning tool.

Data sources

As earlier mentioned, Evac has practically sources of shipboard data, meaning that all more than 30 000 reference ships are disconnected. This means that there is no data infrastructure, no data management methods, no data science nor machine learning tools.

Luckily, this project work coincided with first IoT-piloting project that I am managing as a part of bigger companywide digitalization project. This project was launched in October 2018. It is a first piloting of IoT-system onboard Mein Schiff 2 (MS 2) owned by TUI cruises and operated by Celebrity Cruises and a part of fleet of Royal Caribbean International. In the piloting, the main control panel of the waste management system was connected live to Arnon Sky that is basically built on Azure cloud service.

Because the connection to the ship was not live until late April 2019, data set derived from sistership Mein Schiff 6 (MS 6) was used for demonstrating is data science project work. The data from the MS 2 was later on to validate the machine learning model created with the data from MS 6.

Shipboard data

There are both internal and external data sources that would be used. The internal data sources are the waste data produced from the ship IoT. In the first stage of data exploration, a set of readily produced CSV-files containing all the waste data from Mein Schiff 6 covering couple of days. The raw data set consists of:

- Levels of food waste surface in tank A as percentage of filling. Full tank (100 % filling) equals 5 m3 of volume
- Levels of food waste surface in tank B as percentage of filling. Full tank (100 % filling) equals 5 m3 of volume
- Levels of biosludge waste (BW) surface in the BW tank as percentage of filling. Full tank (100 % filling) equals 10 m3 of volume
- Status information of dry waste (DW) chute valves that are dropping the waste to the waste burning incinerator.

Operational data

To get a sensible picture about the shipboard data, one needs to see the what are the conditions where the waste data is produced. For that purpose, ship operation data was collected from the same time period the shipboard data was collected from following web pages:

Mein Schiff 2:

- Current and future itineraries:
  https://www.cruisetimetables.com/cruise-ship-mein-schiff-2.html
  https://www.cruisemapper.com/ships/Mein-Schiff-2-738

- Past itineraries:
  https://crew-center.com/mein-schiff-2-itinerary

Mein Schiff 6:

- Current and future itineraries:
  https://www.cruisetimetables.com/cruise-ship-mein-schiff-6.html
  https://www.cruisemapper.com/ships/Mein-Schiff-6-1333

- Past itineraries:
  https://www.crew-center.com/mein-schiff-6-itinerary

Before going into exploration of the shipboard data several questions arises:

- Is the production rate constant and/or is the production pattern know? To answer that question one need to know the variables affecting the waste production.

- Activity of passangers that is dependant on the time of a day: That would basically be solved with Pandas datetime-data type.

- Amount of passangers onboard: Basically, it is safe to assume that the ship is always fully booked, that is the situation almost always for all cruise ships in real life. This may be thereby either neglected.

- Proportion of passangers present onboard: Cruise ships visit the ports on a daily basis and most of the cruise ship customers disembark to the port of call. So, the relevant question affecting the waste production; is the ship in a port side or cruising in the sea conditions?

- The amount of food waste is dependant on the meal time, meaning that two hours before and until the meal time the main galleys are fully operational. We can assume the meal times of a day based on our knowledge about the ship operation.

As we can see, there is at least five most relevant attributes affecting the waste production rates. Thereby, one can safely conclude that the production rates aren't constant nor known. We can also conclude that the rates are not same throughout one day. Most likely the production patterns are very ship-specific and cannot be generalized from single attribute. Therefore, the development of the data management and data science tasks is a relevant in this project.

Computational tools

First of all, it is worth of noting that Evac Oy has no background in any kind of programming. So, during this project I practically piloted the usage of data science tools with the real data. This will create foundation for the future development of data science tools in Evac.

The work is done with Jupyter Lab of Anaconda distribution package on a MacOS Mojave 10.14.4. The code was written in Python meaning that all the usable Python libraries were used in the project. Basically, Numpy package [6] is the backbone of all data wrangling on the notebooks running on Python kernel. Numpy works on array-datatypes and is extremely fast as it is constructed with C, Fortran and LVM. Pandas package [7] is built on the top of the Numpy package, but unlike Numpy, it includes indexing rows and columns. It gives handy computational interface for data cleaning, and calculations. The visualisation of cleaned data was done with Matplotlib package [8] and Scikit-Learn [9] for model training and validation.

Data science work

As said, the work was done on Jupyter Lab notebooks with running Python kernel. All the code and data sources were in single folder in on virtual environment. A Github-repository was created for the project (https://github.com/Jarzan/DiplomaInAIProject). Each day, after completing the work, the update was pushed to the repository.

Function as a reusable code-unit

One goal of this work is to create good practice of data science work. This would include creating a library of reusable functions [10]. Basically, all the actions of the data conversions were coded in form of functions. As an example function is shown in Figure 1.

```
def Datetime(df):
    """
    Converts the str-type TimeString into more useful Pandas DateTime form.

    Prints the head of the DataFrame after conversion and the data type of the DateTime column.

    :input: Name of the dataframe and the TimeString column of the DF in question.

    :return: pandas dataframe with the DateTime and its data type.
    """

    df['DateTime'] = pd.to_datetime(df.TimeString, format = '%d.%m.%Y %H:%M:%S')
    dty = type(df.DateTime)    # Returns datatype
    print(df.head())
    print('The time data type is now: ',dty)
    return df, dty
```

Figure 1. Example of a reusable function where the string-type date is converted to Pandas DateTime-datatype.

The function is a reusable code snippet that can be called in the main program by the name of the function, e.g. Datetime (Fig. 1) and the data or variable “treated” by the function is given as parameter inside parenthesis. The first rows fter the name of the function should contain a docstring that is a multiline comment telling what action the function carries out, what are the input parameters and what is the return value of the function. The function can return a value(s) that is stored in a variable(s).

The functions that created in this work are reusable with small modifications in the other Evac shipboard project. Thereby, the data science library is created.

Data import

The data was imported as CSV-files to the notebooks with Pandas read_csv-command. This was applied in form of DataLoad-function, that also returns the dimensions of the data and the data type as well. Owing to the high 5 s sampling frequency the file size of 500 000 rows covered only couple of days of data. Nevertheless, the data of multitude can still be loaded to the Pandas dataframe without any delay.

Evac has no data management system. The shipboard data for the complete delivery will be several millions of data per each day. This will mean that a proper data management system with relational database management system (RDBMS) has to be established. This needs to be done before going into software production phase with the shipboard data.

Data cleaning and filtering

The data was split into several files that included data from various sources of waste management system. The formation of the CSV-files from the process PLC was not logical. The logged data files are formed to the Siemens Simatic PLC to a DB-slots from where the logged data are sent to the cloud server. Basically, the data in the file had no logical order that would have served the use of data. This was solved by picking up each data source into its own dataframe, like e.g. “Food waste tank A” [11]. This was done as Filtvar-function that takes the original dataframe and the desired column as parameters and returns the desired data source as a separate dataframe. This gives possibility filter easily any onboard data source of the process.

In addition, the time stamp to each data point in the logged data, were created in string-data type. This was solved with the Datetime-function as shown in Fig. 1.

All the data sources in the process used 5 seconds sampling frequency as a default. This created extremely noisy data that included every movement of rolling-and-pitching of the ship as well as movements caused by starting and stopping the discharge to the tank. Basically, this noise is and will be there also in all of the ships. It just needs to be acknowledged and omitted.

Non-useful omitted data

During data filtering and exploration, it turned out that part of the planned data use was simple not sensible. The original aim was to use DW data to describe the dry waste production onboard the cruise ship. It turned out that the waste data source described the feeding of the waste into the incinerator and had nothing to do with the actual production of the waste. This can be seen in the timing pattern shown in the Figure 2.

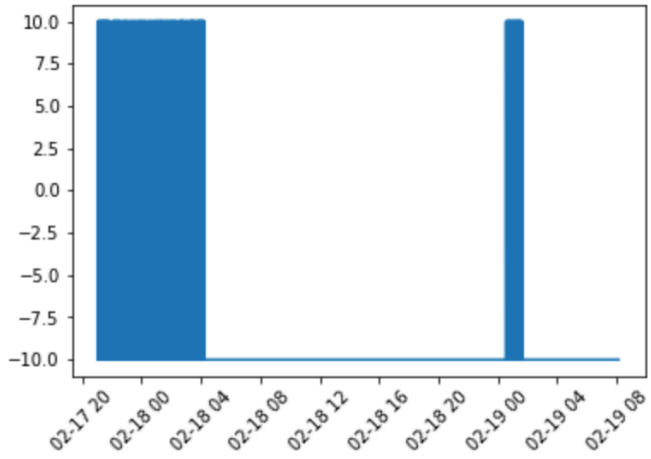

Figure 2. Data visualization for dry waste showing the timing of the waste burning instead of waste formation.

Thereby, the DW as well as the BW data have to be omitted as non-useful data. The DW data will be useful when the data science work is expanded to study the waste burning. The reason for awkward shape of the BW data is not clear. This needs to be studied with the future data. After this decision the work was concentrated to the FW data.

Production rate data

The most desired data information to describe a needed capacity information or pollution emissions is the production rate of individual source. To gain that data from secondary parameter like %-of tank level, several functions needs to be coded. This includes the filtering as described above. In addition the discharge pumping from the tank had to be removed from the dataframe. This was done with a where()-method of the Pandas that basically works like an if-loop, but applying the conditional action only if the condition is False. From the dataframes of individual data source the cumulative waste production was first produced by calculating the difference of each row and thereby calculating

cumulative sum of the differences. The cumulative food waste production is visualized in the Figure 3.

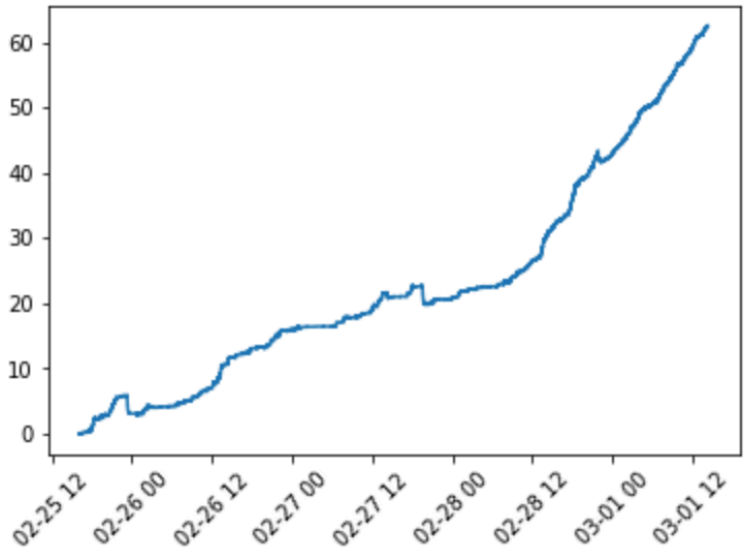

Figure 3. Visualization of the cumulative food waste production in tank 1/A.

The data visualization shows the true nature of filtered data as can be seen in Fig. 3. The cumulative curve is acceptable. This data was used to produce the rate information by calculating the time against the created datetime column. The rate was visualized with a simple .plot()-method of Matplotlib as shown for FW tank 1 in Figure 4.

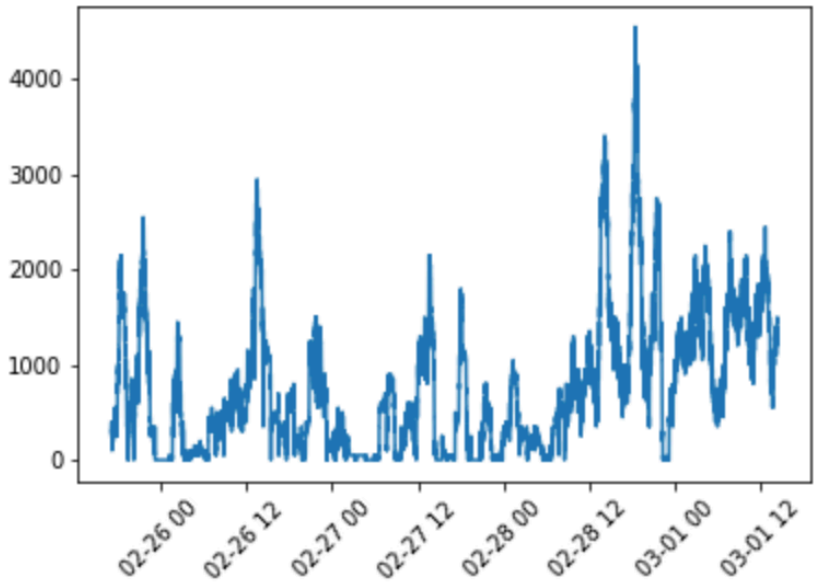

Figure 4. Visualisation of rate of FW production from the ship activities into FW tank 1.

The visualization in Fig. 4 shows how noisy the dense data is. The sampling rate of each process data source has to be considered carefully.

Splitting the port and the sea data

As a final data wrangling task, inPort- and inSea-functions were programmed to split the data into timelines were the ship was in a port and in sea conditions. These functions take the dataframe, and in and out datetimes as parameters and returns a dataframe with the data from the given timeline. In this prototyping work, the datetime was typed in. However, in the production version of this prototype the port and the sea timelines will be webscraped from the web sites given above.

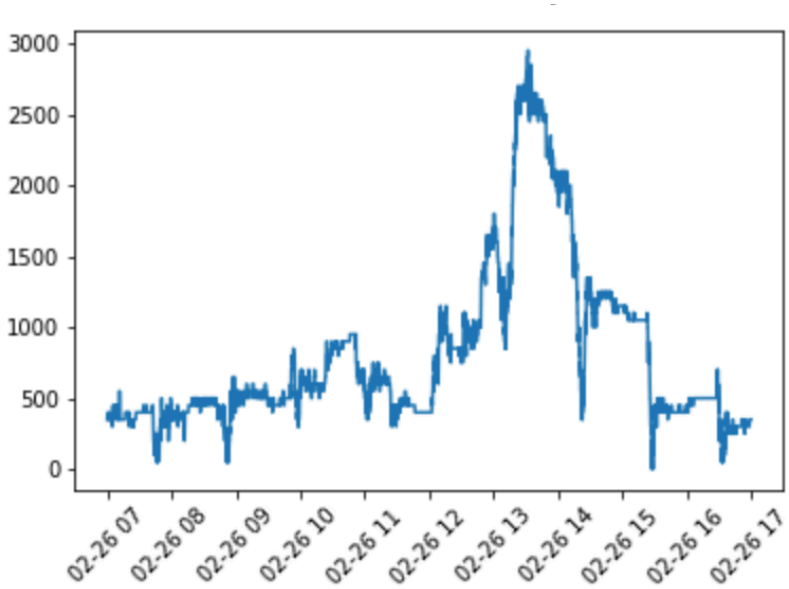

Figure 5. Visualisation of rate of FW production into FW tank 1 during staying in port.

The visualization in Fig. 5 shows the rate of FW production whilst the ship is staying in port from 7:00 o'clock morning unit 17:00 in the afternoon. This shows that the rate of FW production into that tank was highest during afternoon.

Prediction model

The final task of this project is to fit a predictive model to the FW datasets (see the code as appendix). The problem solved by machine learning model is to predict the rate of food waste production. This will help to operate the FW tank storage capacity at each pumping station against the at certain time points whilst staying in port or at sea on certain moment of time.

This task is applied only in an elementary level. Both linear and polynomial regression functionality of the Scikit-Learn-package will be utilised for model development and training the model. In this project, the high-level insight produced by the model is the production of the FW waste streams. From that basis it would be possible for the operators to conclude whether or not the storage capacities are sufficient.

Predicting the rate of FW production from the data

The inputs (X) as relevant features for the ML-model are the time points in the port and the other set of time points are at the sea. The labelled data are the rates of FW production at aforementioned time points in the port and at the sea. When applying the sklearn-models to the data, it turned out that Pandas DateTime-data (X) could not be directly used for the prediction. Therefore, the DateTime-data was converted into float-datatype. The output (Y) predictions will be the predicted waste productions at given time point. Before applying the model the flow rate data in the port was combined.

Both linear [4] and polynomial [5] regression models were constructed from the data. Owing to non-linear nature of the training data, the linear model did not fit well to it. The fitting of the polynomial model to the port data is shown in Fig. 6. The fitting of polynomial model as well as the coding for fitting is shown in Fig. 7. The make_pipeline from sklearn.pipeline-module were used and 4^{th} degree equation were used for the fitted model. Increase of degree as a parameter for PolynomialFeatures()-method did not affect the shape of the model. This is probably due to dense data. The high and low peaks are surely visible but they probably are few of outlier values that don't represent the prevailing level of data. Thereby, the few peaks are not expressed in the curves of the models regardless of the level of polynomial degree.

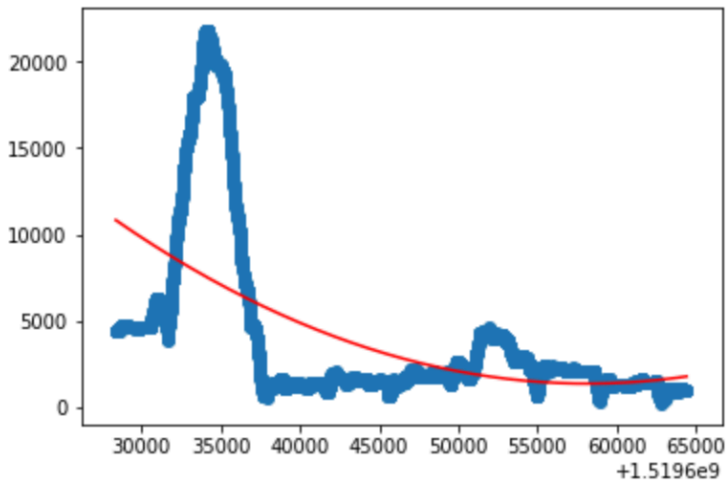

Figure 6. Fitting of polynomial model to the of rate of FW production during staying in port.

Polynomial regression:

[296]:
```
X_poly1S = poly_regP.fit_transform(X1S)
```

[307]:
```
poly_model1S = make_pipeline(PolynomialFeatures(4), LinearRegression())
poly_model1S.fit(X_poly1S, y1S)
y1Sfit = poly_model1S.predict(X_poly1S)
```

[308]:
```
# Plot the sklearn.pipeline-based polynomial model against the data:
plt.plot(X1S,y1Sfit, color='red')
plt.scatter(FW1Sea.FloatingDateTime,FW1Sea.lhFlow)
```

[308]:
```
<matplotlib.collections.PathCollection at 0x1a625e4b00>
```

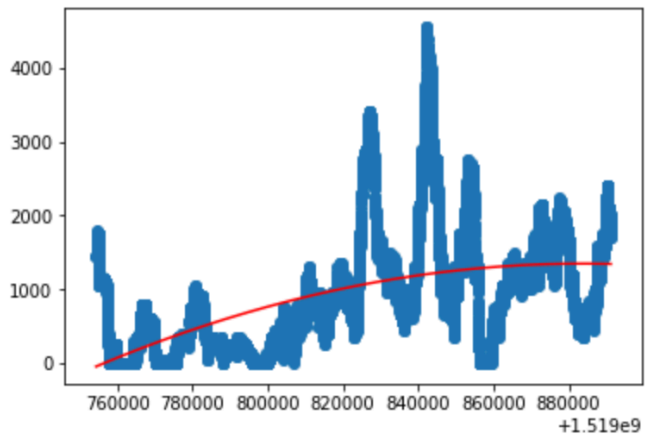

Figure 7. Visualisation of rate of FW production into FW tank 1 during staying in port.

Evaluation of the developed predictive polynomial regression model

A predictive model was produced with a polynomial regression method of Scikit-Learn module [9]. The labeled production rate was modelled against the moments of time whilst being in port or at sea. By the first look, the polynomial model gave fairly good fitting to the data. The error of the model was computed against the training data. The model validation was done with a r2_score method of the Scikit-learn-package. The model did not perform well. More prototyping of the model development needs to be done in order to produce reliable tool for prediction.

Conclusions

The main conclusion of the project work is that the shipboard process data requires lot of data preprocessing to be in any sense useful. This can be minimized by revising the PLC software so that e.g. the datetime is directly useful the variable names are clear and simple and the sampling frequencies per each variable are well thought. Also, the data is derived mainly only as status information that causes multi-stage data transformation before one can make any meaningful visualisation of the data.

Evac has no background in utilizing the data. Firstly, this is seen lack of understanding of meaning of data and its possibilities. On practical level, it is utmost urgent to establish a data management system to have any chances of “surviving” the flood data that will be ahead. There are now several inquiries from customer ships to get the system connected. Building a solid data management system will give good foundation for the digital business and will make the data science and ML operations smoother.

The prediction of the FW production rate against the time points in the port and at sea is a relevant task. This can be done by fitting the polynomial regression model to the labelled data. In this project work the model did not perform well in the validation test. More work is needed to produce useful tool. The validity of the model needs to be studied separately to estimate the reliability and applicability of the produced model. The actual ”acid test” needs to be done with the future shipboard data in a longer term.

The applicability of ML methods needs to be studied separately from the business point of view. A simple prediction task would be fairly easy to incorporate as for example online tool as soon as the onboard process in question is connected and the processing of data is automated. Higher added value data business opportunities need a separate in-depth study.

References:

[1] https://en.wikipedia.org/wiki/Cruise_ship

[2] Cruise ship waste management database. Evac Oy. 2019.

[3] IMO Marpol Annex V, Resolution MEPC. 295(71).
http://www.imo.org/en/OurWork/Environment/PollutionPrevention/Garbage/Pages/Default.aspx

[4] An Introduction to Statistical Learning with Applications in R. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. 2017. eBook. Corrected at 8th printing.

[5] Python Data Science Handbook. Essential Tools For Working With Data. Jake VanderPlas. 2017. 1st Edition. O'Reilly.

[6] NumPy v1.16 Manual, updated 31 Jan 2019.

[7] Pandas manual v0.23.4 Final, updated August 3, 2018.

[8] Matplotlib 3.0.3 documentation.

[9] Scikit-learn manual, 0.20.3 stabile.

[10] Python for Data Analysis. Data Wrangling with Pandas, Numpy and iPython. Wes McKinney. 2018. 2nd Edition. O'Reilly.

[11] PyCon 2018: Using pandas for Better (and Worse) Data Science, Instructor: Kevin Markham, GitHub repository: https://github.com/justmarkham/pycon-2018-tutorial.