IE 48B Project Report

Group 9

Mustafa Kaan Aslan (2017402087) – Yunus Emre Karataş (2020700108) – Ahmet Sualp Say (2020702141)

January 22, 2022

Introduction

In this project, we faced with the classification problem on time series. It was a 2-phased project which consists of trial and live phases. The main purpose was to predict the system net imbalance sign direction until 12pm for the rest of the day before the imbalance occurs in the energy market. In order to train models the imbalance information between January 1st, 2019 and December 10th, 2021 is provided as a bulk *.csv file in trial phase then the data is updated everyday on live phase of the project. There were 1075 days and 25800 hours in total. Additionaly, we were given the weather information for 7 big cities. Preprocess is not needed for the data since all column types looked reasonable and we didn't observe any NA values. The net imbalance volume for all day & hours can be seen below.

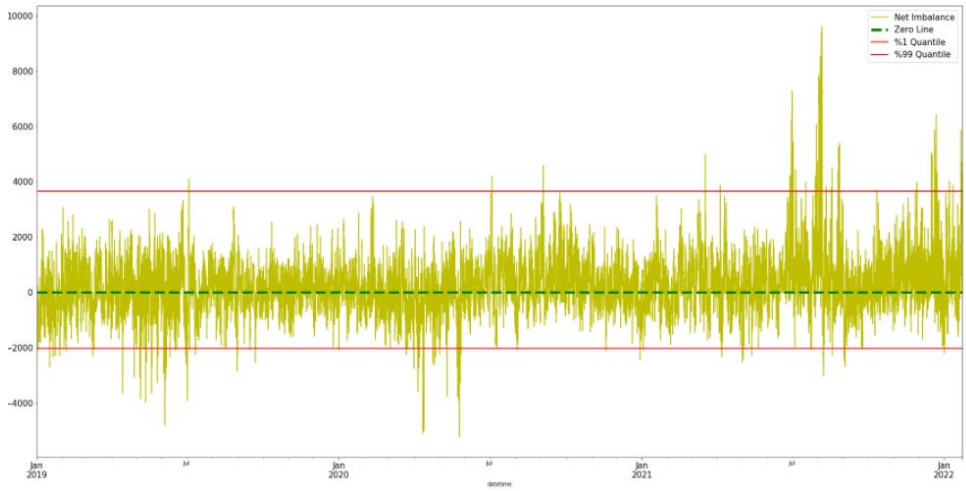

Figure 1: *Net Imbalance Volume*

One can see that there are some outlier periods both in 2020 and 2021. The outlier negative imbalance for the months April, May 2020 may somehow related to pandemic but we were not able to find any information about what causes those huge positive imbalances in the month July and August 2021. Also, other potential outlier points can easily be seen by looking at the box-plot. Day and month averages of imbalance volume can be seen in plots below.

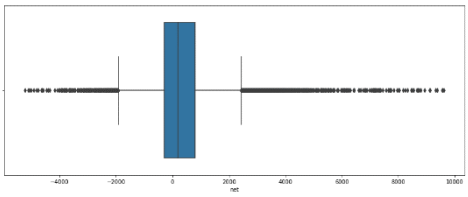

Figure 2: *Box-plot of net imbalance volumes*

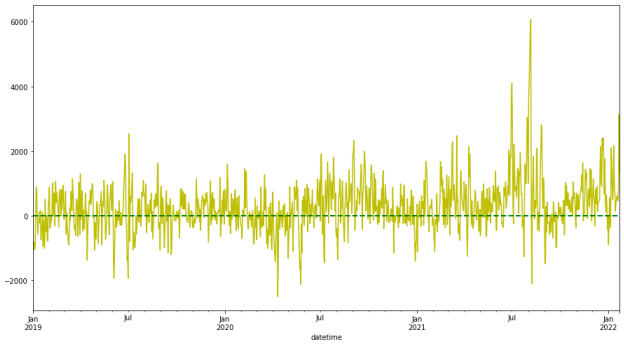

Figure 3: *Day averages*

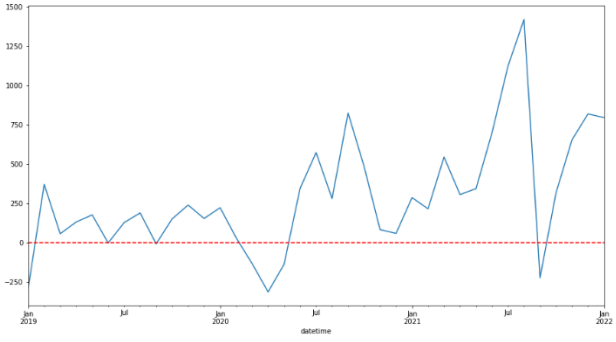

Figure 4: *Month averages*

Since we need to train the model in hourly manner, we plot hours for each day as time series in order to understand the chrachteristics.

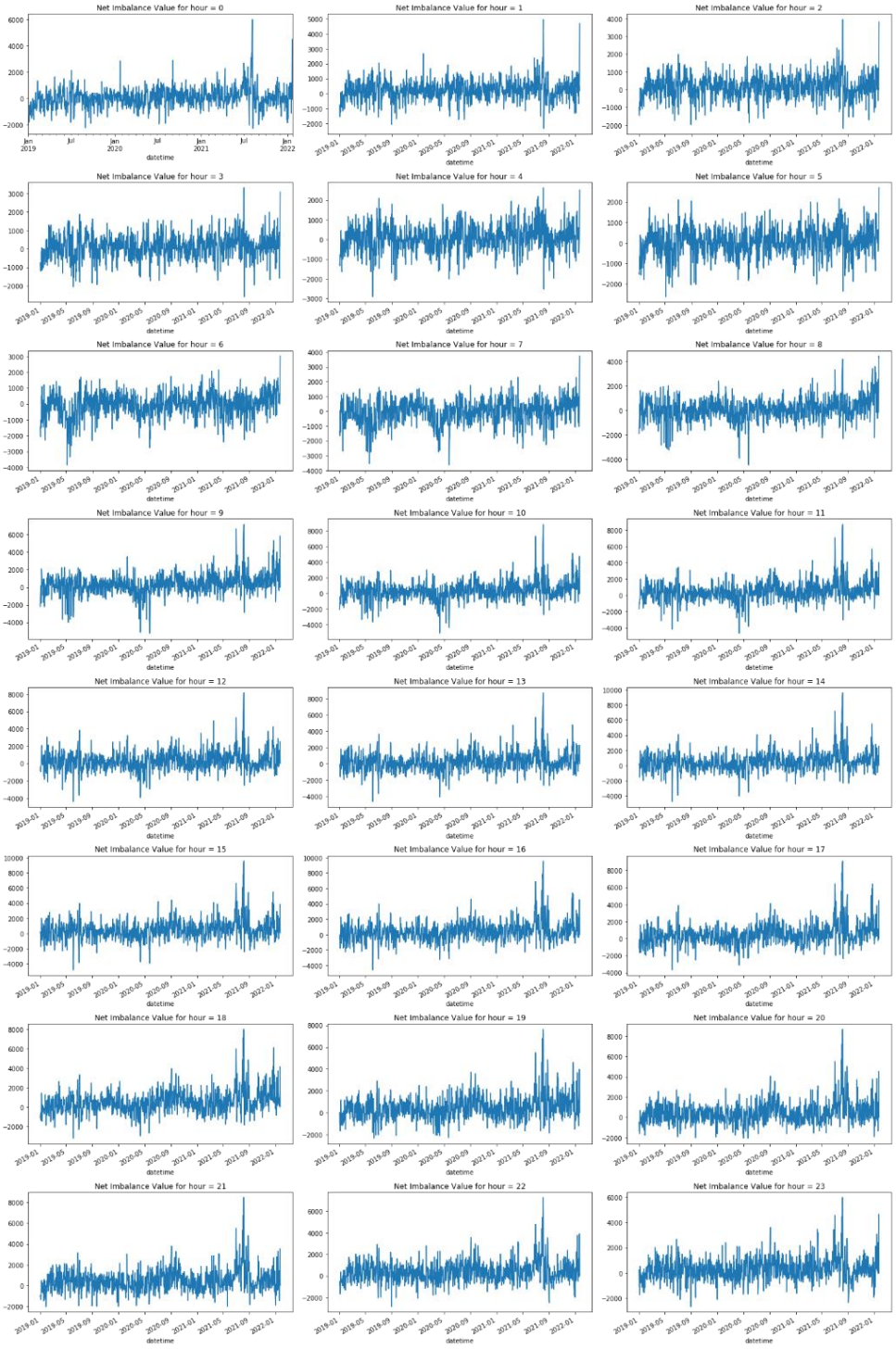

Figure 5: *Hours chracteristics*

We finalize the descriptive analysis part of the project by looking at the distribition of system imbalance sign based on hours, week days and months.

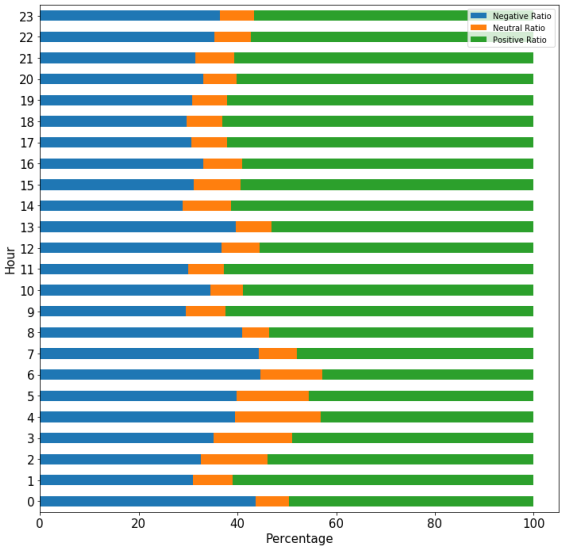

Figure 6: *Hourly Distribution of Classes*

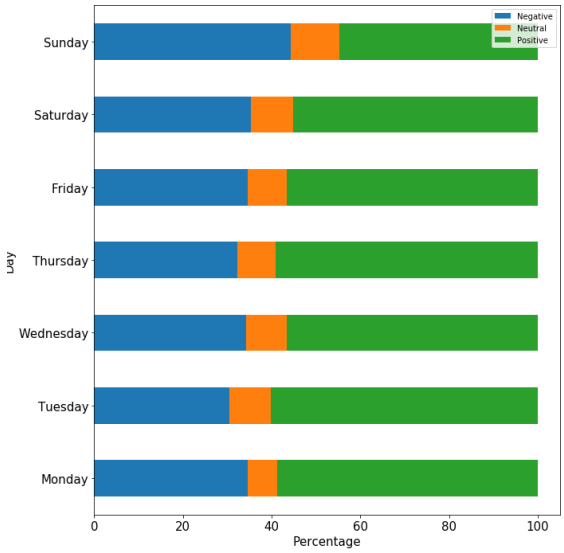

Figure 7: *Distribution of Classes on day of week*

Neutral class contributes to very low proportion in all distribitions and there is a sligtly difference between positive and negative classes.

We proposed different approaches to the problem. Besides distance based and model based approaches, we also tried a combined version. Proposed models are compared with two simple baseline methods which are previous day's same hour's system's direction and previous week's same day same hour's direction. Overall accuracy comparison for all models in live prediction period are summarised in the table below. Details and accuracy of models for each hour-day combination can be found in the results part.

Model	**Accuracy**
Baseline 1 (lag24)	69.05%
Baseline 2 (lag168)	69.05%
Selected Model	74.40%

Table 1: *Overall accuracy for live period*

Related Literature

Baydogan et al. [1] proposes a method that can use all features in multivariate dataset simultaneously. They benefit from tree classifiers that uses time input, difference (for trend information) and additional attributes together to find the regions that one class could dominate. The representations are the terminal nodes of the tree and random forest classifier can find complex regions since each tree uses different portion of the data and attributes. Then, the representations are used to map time series into the high-dimensional regions (say 5 regions) that one class can dominate. Later, BoW method is used to find normalized frequency of time points in each region for a time series. Finally, the BoW features are used to classify each time series correctly.

Approach

Since the weather information is very important to the electricity consumption, we try to take advantage of the weather forecast data in the project. In addition to the raw time series, two methods are tried to get univariate weather time series data. First, PCA is adopted and the first principal component that covers 84% of the variance is extracted as time series sequence. Moreover, random forest model is constructed using 2019 and 2020 data. The fitted RF model is then used to produce time series sequence starting from January 2021. As a result, PCA transformed and RF transformed raw sequences are considered as the input for the next steps

Imbalance prediction for electricity market task is a time series classification problem and thus raw input should be converted into time series that previous time steps are considered as features. In this project, we choose to use 12,24 and 36 timestep windows. After the data preparation is completed, cross validation splits are constructed and saved in the files for reproducibility. To overcome the noise problem and better capture the temporal characteristic of the time series, decision tree and SAX representation techniques are adopted in addition to the raw input. For each 2 raw input options we apply representations and so, 6 possible time series are obtained before the distance calculation

Euclidian distance is evaluated as a base metric. Besides, DTW and ERP metrics are used. DTW deals with the problem of scaling variation and dilation and thus Sakoe-Chiba DTW with its own window constraint is evaluated. ERP is considered as more robust to noise compared to DTW, so it is also investigated. Distance matrices are constructed before the competition and they are reused to find the imbalance class of the prediction day input.

Number of neighbors in KNN should be carefully determined since too high or low number could mislead the prediction.Therefore, options in different scale such as 10, 20, 30 and 40 are evaluated separetely. In the end, the proposed method is compared with the baseline methods (class of 24, 168 hour before). The mean and standary deviation of the accuracy show that our proposed method is better than the baseline methods. Therefore, overall pipeline that performs the explained steps is used in every day predictions.

Detailed version of the approach is embeded to code and can be accessed through the link provided below.

Results

We obtained promising results using selection strategy for each hour mentioned above. Daily results of proposed model and baseline models can be seen in the Table2. Proposed model's overall accuracy in the test period is 0.744 which is higher compared to both Baseline1 and Baseline2.

date	Baseline1_Daily_Accuracy	Baseline2_Daily_Accuracy	Model_Daily_Accuracy
<chr>	<dbl>	<dbl>	<dbl>
2022-01-09	0.7500000	0.5000000	0.7500000
2022-01-10	0.5000000	0.8333333	0.6666667
2022-01-11	0.6666667	0.5000000	0.9166667
2022-01-12	1.0000000	0.3333333	1.0000000
2022-01-13	0.8333333	0.8333333	0.7500000
2022-01-14	0.4166667	0.3333333	0.4166667
2022-01-15	0.2500000	0.9166667	0.6666667
2022-01-16	0.6666667	0.8333333	0.6666667
2022-01-17	0.1666667	0.5833333	0.3333333
2022-01-18	0.4166667	1.0000000	0.9166667
2022-01-19	1.0000000	1.0000000	0.9166667
2022-01-20	1.0000000	0.8333333	0.7500000
2022-01-21	1.0000000	0.2500000	0.7500000
2022-01-22	1.0000000	0.9166667	0.9166667
Mean	0.6904762	0.6904762	0.7440476

Table 2: *Daily results of proposed and baseline models*

Besides aggregated summary, we also wanted to check hourly accuracy for live period since our model trained in hourly manner to see whether there were any spesific hour it consistently makes mistake.

hour	Baseline1_Hourly_Accuracy	Baseline2_Hourly_Accuracy	Model_Hourly_Accuracy
<int>	<dbl>	<dbl>	<dbl>
12	0.5000000	0.3571429	0.2142857
13	0.5714286	0.5714286	0.7142857
14	0.5714286	0.7857143	0.7142857
15	0.8571429	0.7857143	0.7142857
16	0.6428571	0.4285714	0.6428571
17	0.8571429	0.8571429	0.9285714
18	1.0000000	0.8571429	1.0000000
19	0.5714286	0.6428571	0.7857143
20	0.5714286	0.6428571	0.7857143
21	0.5714286	0.7142857	0.7857143
22	0.7142857	0.7857143	0.8571429
23	0.8571429	0.8571429	0.7857143

Table 3: *Hourly results of proposed and baseline models*

One can see from the hourly results that there might be something special about Hour12. Even baseline models are not good at predicting it, our model is doing slightly worse than that. This means, there is a room to improve the model by learning from its errors.

Conclusions and Future Work

We were asked to construct a model predicting system imbalance sign for energy market. It was not an easy task due to requirement of domain knowledge. We achived 5% decrease in the error rate compared to baseline models just by using the limited source of time series representation techniques and computational power. To go further, one can

- Use other external data sources except for the weather
- Check for what's happenning about Hour12 regularly
- Try alternative representation tecniques like Longest Commen Subsequnces or parameter settings for already used ones
- Try broader window sizes when constructing representations
- Leverage different classifiers or ensemble methods instead of just kNN

Code

Descriptive Analysis (ipynb)

Approach, Results and Model Selection (ipynb)

Descriptive Analysis (html)

Approach, Results and Model Selection (html)

References

Baydogan, M.G., Runger, G. Learning a symbolic representation for multivariate time series classification. Data Min Knowl Disc 29, 400–422 (2015). https://doi.org/10.1007/s10618-014-0349-y