TASK 2

POWER PRICE PREDICTION

You are in the group Deepforgetting consisting of junterhol (junterhol@student.ethz.ch (mailto://junterhol@student.ethz.ch)), merklec (merklec@student.ethz.ch (mailto://merklec@student.ethz.ch)) and ribadov (ribadov@student.ethz.ch (mailto://ribadov@student.ethz.ch)).

1. READ THE TASK DESCRIPTION

2. SUBMIT SOLUTIONS

3. HAND IN FINAL SOLUTION

1. TASK DESCRIPTION

INTRODUCTION

Applying ML approaches to real-world problems requires some extra steps in handling specific data artifacts. Some common challenges you may face are missing values, an imbalance of the labels, or noise in the data distribution. These can be addressed through a specific choice of pre-processing and/or model components.

In this task, as an illustration of a real-world problem, you are asked to predict the electricity prices in Switzerland given price information of some other countries and additional features. You will encounter typical ML workflow challenges of missing features and low predictivity in this task.

The following sections provide more details on the dataset, submission and evaluation.

DATA DESCRIPTION

Download handout (/static/task2_k49am2lqi.zip)

In the handout for this project, you will find the the following files:

- **train.csv** - the training set
- **test.csv** - the test set file to make predictions on
- **sample.csv** - a sample submission file in the correct format

- **template_solution.py** - a template file that will guide you through the implementation of the solution
- **template_solution.ipynb** - a template file in jupyter notebook format that will guide you through the implementation of the solution

You are free to use either jupyter notebook or the .py template file.

Each line in **train.csv** represents one data point and consists of the following structure:

```
season,price_AUS,price_CHF,price_CZE,price_GER,price_ESP,price_FRA,price_UK,price_
spring,,9.644027877268496,-1.6862480951361345,-1.7480763846576997,-3.666005401185
...
```

As mentioned in the introduction, we are interested in predicting the electricity (log) prices in Switzerland (corresponds to price_CHF column), given the prices in some other countries (corresponds to all columns starting with "price_" except price_CHF), and additional features (corresponds to season). Note that there might be missing values in the price_CHF column as well.

test.csv is the file you submit predictions on. The file consists of the following structure:

```
season,price_AUS,price_CZE,price_GER,price_ESP,price_FRA,price_UK,price_ITA,price_
spring,,0.4729846640395274,0.7079565159369821,,-1.1364407652408537,-0.59670285573
...
```

For your convenience, a sample solution file **sample.csv** is provided with the following structure:

```
price_CHF
0.635389524381449
0.635389524381449
0.635389524381449
0.635389524381449
```

template_solution.py provides a starting template structure for how you can solve the task, by filling in the TODOs in the skeleton code. It is not mandatory to use this solution template but it is recommended since it should make getting started on the task easier.

MODELLING TIPS

DATA IMPUTATION

Missing data is a commonly encountered artifact in several machine learning tasks. Typical imputation strategies to deal with missing data include:

- Discarding rows or columns with missing data
- Replacing missing values with the corresponding mean or median
- Advanced imputation strategies based on iterative model fitting.

Additional resources on data imputation can be found in this kaggle notebook (https://www.kaggle.com/code/residentmario/simple-techniques-for-missing-data-imputation/notebook) or sklearn website (https://scikit-learn.org/stable/modules/impute.html).

HANDLING CATEGORICAL (NON-NUMERIC) DATA

Some data in this task is categorical (non-numeric). This is a common challenge in machine learning. Some strategies on how to handle categorical data can be found in this kaggle notebook (https://www.kaggle.com/alexisbcook/categorical-variables) or sklearn website (https://scikit-learn.org/stable/modules/preprocessing.html#encoding-categorical-features).

KERNELIZED REGRESSION MODELS

You saw kernelized estimators in the lecture notes. In this task you might find them useful. The core of the challenge is to pick the right kernel for the regression. Commonly used kernels are the linear (or dot product) kernel, squared exponential (or RBF) kernel, polynomial, Matern, and RationalQuadratic kernels among many others. A probabilistic (Bayesian) equivalent of kernelized ridge regression goes by the name Gaussian processes. It provides a principled modelling paradigm with uncertainty estimates. The uncertainty component is not important for this task, however a lot of machine learning software packages implement this method in a very efficient manner and its mean prediction does the same as kernelized ridge regression. The following code block gets you started with gaussian processes in sklearn (more resources can be found here (https://scikit-learn.org/stable/modules/classes.html#module-sklearn.gaussian_process))

```
from sklearn.gaussian_process import GaussianProcessRegressor
from sklearn.gaussian_process.kernels import DotProduct, RBF, Matern, RationalQuadratic
gpr = GaussianProcessRegressor(kernel=DotProduct())
gpr.fit(X_train, y_train)
```

Finding the right kernel can be done in multiple ways as you saw in the lectures:

- Using a validation set
- Cross-validation
- Maximizing the evidence of a Bayesian model. In this case, we are maximizing the total probability of the data given its generative model. This is also referred to as Bayesian model selection. More information can be found here: Bayesian model selection - lecture notes. (https://www.cse.wustl.edu/~garnett/cse515t/fall_2019/files/lecture_notes/7.pdf)

All the above methods are implemented in scikit-learn. Note that Gaussian processes implement Bayesian model selection automatically.

SUBMISSION FORMAT

The submission file format should be the same as that of **sample.csv**, i.e. the file must contain a column with name price_CHF, and each row is the corresponding prediction. Please ensure that the number of predictions in your submitted file are the same as number of data points in the **test.csv** file.

Furthermore, please keep in mind that as a group, you have a limited number of submissions as stated on the submissions page.

EVALUATION

We are interested in an accurate prediction of electricity prices in Switzerland. As mentioned before, this corresponds to the price_CHF column in your submitted predictions. Your submitted predictions will be evaluated by the *R-squared* (R^2) metric to the true prices (here, a higher score is better).

$$R^2(\mathbf{y}^*, \mathbf{y}) = 1 - \frac{\sum_{i=1}^{N} (y_i^* - y_i)^2}{\sum_{i=1}^{N} (y_i^* - \bar{y})^2}$$

where $\bar{y}$ denotes the average of true values $\{y_i^*\}$.
To calculate the R^2 score, we use the scikit-learn implementation:

```
from sklearn.metrics import r2_score
r2_score(y_true, y_pred, squared=False)
```

GRADING

We provide you with **one test set** for which you have to compute predictions. We have partitioned this test set into two parts (of the same size) and use it to compute a *public* and a *private* score for each submission. You only receive feedback about your performance on the public part in the form of the public score, while the private leaderboard remains secret. The purpose of this division is to prevent overfitting to the public score. Your model should generalize well to the private part of the test set. When handing in the task, you need to select which of your submissions will get graded and provide a short description of your approach. This has to be done **individually by each member** of the team. We will then compare your selected submission to our baselines. This project task is graded with grades between **2.0 - 6.0**. Your grade is calculated using a weighted sum of your private and public score. We do not share the weights used for the grading, again to prevent overfitting to the public score. The weights are selected such that:

- To achieve the best grade (6.0), you need to perform better than the hard baseline in both private and public score.
- To pass the project (grade: 4.0), you need to perform better than the easy baseline in both private and public score.

The medium baseline is only used for your reference and is not considered in the grading. In addition, for the grading, we consider the code and the description of your solution that you submitted. The following **non-binding** guidance provides you with an idea on what is expected to pass the project: If you hand in a properly-written description, your source code is runnable and reproduces your predictions, and your submission performs better than the baselines, you can expect to have passed the assignment.

⚠ Make sure that you properly hand in the task, otherwise you may obtain zero points for this task.

PLAGIARISM

The use of open-source libraries is allowed and encouraged. However, we do not allow copying the work of other groups / students outside the group (including work produced by students in previous versions of this course). Publishing project solutions online is not allowed and use of solutions from previous years in any capacity is considered plagiarism. Among the code and the reports, including those of previous years, we search for similar solutions / reports in order to detect plagiarism. Use of GPT3 Copilot or similar code/language generation tools in any capacity for writing code or reports will be considered and treated as plagiarism in the context of this course. Basic code autocompletion such as those used in the default setup of Sublime Text 3 are permitted. If we find strong evidence for plagiarism, we reserve the right to let the respective students or the entire group fail in the IML 2024 course and take further disciplinary actions. By submitting the solution, you agree to abide by the plagirism guidelines of IML 2024.

FREQUENTLY ASKED QUESTIONS

⦿ WHICH PROGRAMMING LANGUAGE AM I SUPPOSED TO USE? WHAT TOOLS AM I ALLOWED TO USE?

You are free to choose any programming language and use any software library. However, **we strongly encourage you to use Python**. You can use publicly available code, but you should specify the source as a comment in your code.

⦿ WHAT TO DO IF I CAN'T RUN THE CODE/SETUP AN ENVIRONMENT ON MY PC?

If you are having trouble running your solution locally, consider using the ETH Euler cluster to run your solution. Please follow the Euler guide (/static/euler-guide.md). The setup time of using the cluster means that this option is only worth doing if you really can't run your solution locally.

⦿ AM I ALLOWED TO USE MODELS THAT WERE NOT TAUGHT IN THE CLASS?

Yes. Nevertheless, the baselines were designed to be solvable based on the material taught in the class up to the second week of each task.

⦿ IN WHAT FORMAT SHOULD I SUBMIT THE CODE?

You can submit it as a single file (main.py, main.ipynb, etc.; you can compress multiple files into a .zip) having max. size of 1 MB. If you submit a zip, please make sure to name your main file as *main.py* (possibly with other extension corresponding to your chosen programming language, e.g. .ipynb).

⦿ IN WHAT FORMAT SHOULD I SUBMIT THE REPORT?

The handin page of the submission server contains a simple textbox in which you should insert your report. It should consist of a couple of sentences explaining the main ideas and concepts of your solution. Every student writes and submits the report independently.

⦿ WILL YOU CHECK / RUN MY CODE?

We will check your code and compare it with other submissions. We also reserve the right to run your code. Please make sure that your code is runnable and your predictions are reproducible (fix the random seeds, etc.). Provide a readme if necessary (e.g., for installing additional libraries).

⦿ SHOULD I INCLUDE THE DATA IN THE SUBMISSION?

No. You can assume the data will be available under the path that you specify in your code.

⦿ CAN YOU HELP ME SOLVE THE TASK? CAN YOU GIVE ME A HINT?

As the tasks are a graded part of the class, **we cannot help you solve them**. However, feel free to ask general questions about the course material during or after the exercise sessions.

⦿ CAN YOU GIVE ME A DEADLINE EXTENSION?

> ⚠ We do not grant any deadline extensions!

⦿ CAN I POST ON MOODLE AS SOON AS I HAVE A QUESTION?

This is highly discouraged. Remember that collaboration with other teams is prohibited. Instead,

- Read the details of the task thoroughly.
- Review the frequently asked questions.
- If there is another team that solved the task, spend more time thinking.
- Discuss it with your team-mates.

⦿ WHEN WILL I RECEIVE THE PRIVATE SCORES? AND THE PROJECT GRADES?

We will publish the private scores, and corresponding grades before the exam the latest.