KIVOS

Challenge – Data Scientist

The target of the following task is to make clear to us:

- your way of thinking
- your approach to Machine Learning/Data Science problems
- your ability to write quality code
- your ability to express yourself writing a document / your writing skills

Demand forecasting is an essential and demanding task that a retail company has to deal with, in order to optimize its supply chain. In this task, you have to model the demand forecasting problem, as a supervised machine learning problem and predict the future demand for all the item-store combinations in the dataset (described below), for each day in the period from **12th Sept 2022 to 18th Sept 2022**. The assumption is that all the stores operate every day and there are no stock outs.

In particular, the basic steps you have to follow are:

1. **Data loading**

   You will work with three CSV files: sales.csv, regular_price.csv, and promo_price.csv. These files contain historical sales, regular prices, and promotional prices for some item-store combinations from a retail company. You don't have records in promo_price.csv, for the days when there is no promotion activity.

2. **Cleansing**

   For this step, you are required to develop a data cleansing approach. Specifically, you should implement techniques to identify and handle possible outliers in the data. Define what could be considered an outlier based on the context of the sales, and explain your approach for dealing with these outliers (e.g., removal, transformation, or imputation). Your approach should be applied to the data from the provided CSV files, and you should justify the techniques you choose.

3. **Preprocessing and feature engineering**

   This is the most important step of every machine learning pipeline. Try to find useful features (maximum 10-15) and describe them. Also, do the required preprocessing so that the data are ready to be fed in the next step's models. Remember to write a few suggestions about any additional features that could be used to improve the forecast quality, in case you had access to more data.

4. **Selection of models, training, prediction and evaluation**

   In this step you have to select models (maximum 3), train them, and generate predictions. Also, you should describe in short some of your trials and the general process you followed in order to assess the models (training set, test set, etc.). Finally, you should conclude what was the best model for the specific set of data. Please notice that the evaluation of the forecast quality is quite a complex issue. Please, do some brief research on this topic and use some of the most common measures. Don't expect to find the perfect solution.

   Reminder: please do not conduct exhaustive experiments, we just want to be aware of your ability to do the basics.

5. **Saving the results**

   In the last step, you should create a separate CSV file for each model you experimented with and save the predictions generated from each model. For example, if one of the models used is called 'zzz', you should create a file named forecast_zzz.csv and save the forecasts generated by this model. Each CSV file should contain four columns: "date", "item", "store", and "prediction" (e.g., 20220912, item1, store1, 2.25)

The development of this task must be done using any programming language (preferably Python 3).

Deliverables

Your submission should be in a .zip file with following contents:

- the source code
- the CSV files with the predictions
- a readme file with instructions on how to run the code. Please mention possible external required packages
- a document where you describe your approach to the problem's solution and your answers to all the questions above (maximum 3-4 pages)

For any questions or concerns for the task, feel free to send an email to aris.moustakas@kivos.ai