# Submission: 25-Aug 2:45am

# Willow Cash Flow Prediction

Challenge Overview

The goal of this challenge is to forecast Working Capital which is made up of 4 elements

1. Inventory
2. Receivables
3. Payables and Accruals
4. Returns and Rebates

These elements are forecasted at market-level and then rolled up to region.

Additionally, the following are provided as there is expected to be some relationship between these and the working capital elements. Although some guidance is provided below, the exact nature of the relationship is unclear, and will need to be discovered as part of the model.

* Net Sales: The Net Sales is expected to be related to Receivables and "Returns and Rebates"
* Gross Profit: The Cost of Sales (Difference between Net Sales and Gross Profit) is expected to be related to Inventory and "Payables and Accruals"
* Advertising and Promotions: The A&P is expected to be related to "Payables and Accruals"
* Operating Profit: The difference between Gross Profit and Operating Profit, and then removing A&P, is "Other Expenses". It is expected that "Other Expenses" is related to "Payables and Accruals", although this relationship will be different from that of A&P above.

Note that the above ones are cumulative each year. To get the numbers for that period, you will need to take the difference between that period and the previous period (in the same year). For example, Y2-M2 numbers are for Year 2 Months 1 & 2. To get the Year 2 Month 3 number, you will have to take the difference between Y2-M3 and Y2-M2. To get the Year 2 Month 4 number, you will have to take the difference between Y2-M4 and Y2-M3.

Please note that it is not mandatory to use all the data in your model. Also, NDA is required for this challenge --- By registering this challenge, you have agreed to the Topcoder’s NDA.

Challenge Data

We will provide two time series (i.e., “Version” in the data):

1. ACT: It means “Actual”. These are the actual values.
2. PLN: It means “Plan”. These are the PLNs created for each market. Markets aim to achieve the PLN number at the end of each year (number in month 12).

All these values have already been adjusted to Constant Exchange Rates.

Here are some notes for the dataset:

* There may be some structural breaks in the model, for example, due to acquisitions.
* For each year, month 1 is missing. This is a characteristic of this data set - the first observation is 2 months of data. Each year has 11 observations (M02 - M12).
* PLN may not have numbers every month - this is a characteristic of this data set.

Overall, we have 5 years of ACT data (Y1 - Y5) + Y6 M02 and 6 years of PLN data (Y0 - Y5) , and for year 6, we have up to M06 - is provided. The goal of this challenge is to forecast the 4 ACTUAL Working Capital elements for Y6 M03 to Y6 M06.

We have prepared the two csv files (available in the forum):

1. training\_data.csv has the training data that you can use. Please also try to find some way to tune your hyperparameters based on this training set.
2. empty\_test\_sheet.csv is the one that your model must fill. Please rename it as “test\_sheet.csv” when you make the submission.

Some definitions:

* Roll ups: Total Working Capital = Inventory + Receivables + Payables and Accruals + Returns and Rebates
* Markets roll up into Regions as follows: Region A: Market A1 + Market A2 + …..   Region B: Market B1 + Market B2 + …

To re-emphasise, it is not mandatory to use all the data in the model. It may be that a simple model based on Working Capital ACT data is most effective.

Dataset Anonymization

‘Z-score’ is used to anonymize the real data. A variable can be uniquely identified by the triplet <Account-Mapping, Version, Market-Mapping>. For all the variables, following is the formula used to privatise the data:

            zi = (xi – μ) / σ

where zi = z-score of the ith value for the given variable

            xi  = actual value

            μ = mean of the given variable

            σ = standard deviation for the given variable

**Final Submission Guidelines**

Submission Format

Your submission must include the following items

* The filled “test\_sheet.csv” file.
* A report about your model, including data analysis, model details, local train/test splits, k-fold cross validation results, and variable importance.
* A deployment instruction about how to install required libs and how to run.

Expected in Submission

1. Working Python code (could be Jupyter-Notebooks) which works on the different sets of data in the same format
2. Report with clear explanation of all the steps taken to solve the challenge and how to run the code
3. No hardcoding (e.g., hyper-parameters in your model must be adaptive to the training set) in the code is allowed. We will run the code on some different datasets
4. All models in one code with clear inline comments
5. Flexibility to extend the code to forecast for additional months

Quantitative Scoring

Given two values, one ground truth value (gt) and one predicted value (pred), we define the relative error as:

            MAPE(gt, pred) = |gt - pred| / |gt|

We then compute the raw\_score(gt, pred) as

            raw\_score(gt, pred) = max{ 0, 1 - MAPE(gt, pred) }

That is, if the relative error exceeds 100%, you will receive a zero score in this case.

We will evaluate your predictions at different levels:

1. Market-level: For all your forecasts in the filled “test\_sheet.csv”, we will calculate the raw scores against the ground-truth values.
2. Region-Level: For each element (e.g., Inventory), each region (e.g., A), and each month (e.g., Y6-M03), we will first sum up the values in all its markets (e.g., A01, A03, ...) from both your predictions and the ground-truth values, and then calculate the raw score.
3. Element-Level: For each element (e.g., Inventory) and each month (e.g., Y6-M03), we will first sum up the values in all markets (e.g., A01, A03, B01, ...) from both your predictions and the ground-truth values, and then calculate the raw score.
4. Total Working Capital Level: For each month (e.g., Y6-M03), we will first sum up the values in all markets (e.g., A01, A03, B01, ...) and all 4 elements (e.g., Inventory) from both your predictions and the ground-truth values, and then calculate the raw score.

The final score will be an average of all these levels of raw scores, and then multiplied by 100.

Final score = 100 \* average( all raw scores )

More details can be found in the *scorer.py* script (available in the forum).

Judging Criteria

Your solution will be evaluated in a hybrid of quantitative and qualitative way.

* Effectiveness (80%)
  + We will evaluate your forecasts by comparing it to the ground truth data. Please check the “Quantitative Scoring” section for details.
  + The higher final score (i.e., smaller MAPE) the better.
* Clarity (10%)
  + The model is clearly described, with reasonable justifications about the choice.
* Reproducibility (10%)
  + The results must be reproducible. We understand that there might be some randomness for ML models, but please try your best to keep the results the same or at least similar across different runs.