Using Naïve Models to Improve US Dollar Exchange Rate Trend Prediction

Elia Yathie Matsumoto
EESP/FGV
Sao Paulo-SP-Brazil

Emilio Del-Moral-Hernandez
POLI/USP
Sao Paulo-SP-Brazil

Claudia Emiko Yoshinaga
EAESP/FGV
Sao Paulo-SP-Brasil

Afonso de Campos Pinto
EESP/SP
Sao Paulo-SP-Brazil

ABSTRACT
This paper extends our previous research, which proposed a methodology based on the following hypothesis: dealing with the problem of predicting the next-day USD/BRL exchange rate daily trend, the existence of calendar effects allows us to improve trained voting-based ensemble models without model retraining. In the present work, we propose adding naïve models to the originally proposed methodology because naïve models would also potentially benefit from the calendar effect becoming a benchmark to consider. The experiments confirmed that naïve models are not just challenging benchmarks but also models that can be included in the process to improve existing voting-ensemble models. On average, adding the naïve models to the original solution generated an increase higher than 100% in the value of the primary metric adopted for performance measurement. Constantly overwhelmed by more complex solutions, we can take these outcomes as a reminder not to neglect simplicity.

General Terms
Financial Time Series Forecasting, Machine Learning, Behavioral Finance.

Keywords
USD/BRL Exchange Rate, Calendar Effect, Ensemble Models, Naïve Model.

1. INTRODUCTION
In recent years, we have witnessed a boost in research proposing new computational models based on increasingly larger and more complex architecture. Moreover, some huge models, like large language models have delivered outstanding positive results[3, 10]. All of this may convey a notion opposed to the classic "Occam's Razor" principle, which essentially states that "the simpler is better"[2]. Also, the latter statement carries the sense of competition between simpler and more complex methods. In this perspective, conversely, this present work proposes exploiting possible uses of simpler models to improve complex models instead of competing with them by improving our previous research[5], in which we applied a combination of techniques to address a financial time series forecasting problem.

Specifically, we focused on predicting the trend of the daily quotations for the US dollar to the Brazilian real exchange rate (USD/BRL). This kind of task is still challenging, as researchers worldwide continue struggling to get reasonable results even when using larger models consistently[1, 8]. Technical literature attributes this hardship to the complex nature of the economic and financial phenomena[4]. Thus, even many decades after the famous "The Meese-Rogoff puzzle" article publication on pricing forecasting[6], several papers are still adopting random-walk models as a typical benchmark[7, 11].

Our previous article described "a method to improve existing voting-based ensemble models trained to predict the next-day USD/BRL exchange rate trend with no need for retraining or other costly computational tasks.". Despite achieving promising results, some aspects of the performed experiments made us question the robustness of the proposed method during the pandemic and whether it could overcome the simplest random model possible, the naïve model, which takes the last observed value as the prediction for the next.

In the current paper, we describe what we did to answer these questions: (1) we repeated the original experiments, just using more recent data to include the pandemic period; (2) we added the comparison with the naïve models' outcomes and verified that this model category produced the best final result, surpassing all the others; (3) finally, we added the naïve models to the ensemble models and this final combination provided very satisfactory results.

2. RELATED WORK HIGHLIGHTS
The novelty we brought in our previous research [4] was adding a Behavioral Finance element, the calendar effect, to construct a strategy to improve trained voting-based ensemble models without retraining. It only presumes the availability of the past predictions produced by the trained voting-based ensemble model and all the individual models that compose it.

Assuming the calendar effects affect investors' actions, the rationale is that the collective actions of investors constantly trying to raise their trading positions would induce periodic and deterministic patterns in the financial time-series movement. Accordingly, the forecasting models built to predict these financial time-series trends would also be prone to induce periodic and deterministic patterns in their outcomes, allowing us to identify the 'best model' among all available models based on their past performance. Hence, to improve the voting-based ensemble model outcomes, the study proposed replacing its original voting prediction with the one produced by the model identified as 'best' for each trading day.

Since months have different numbers of days, the regular date numbering was not convenient to compare daily metrics on a month-by-month basis. Thus, to better evaluate the models' past performance, each trading day received a 'tag' based on its relative position to the month's first and last trading days. This 'tag' index, composed of 22 values, grouped the trading days into 22 groups, allowing us to compare the models' past performance more uniformly. The model composed of the outcomes of the 'best model' among all available models was named the' Best Model per Tag' ensemble model, the BMT.

We applied the methodology in experiments using 15 years of USD/BRL observations, a total of 3,756 values from July 1, 2005, to June 30, 2020. The voting-based ensemble model (VOTING) was composed of five model categories: K-Nearest-Neighbor (KNN), Logistic Regression (LOGREG), MLP (Multi-Layered Perceptron Neural Networks), RF (Random Forest), and Support Vector Machine (SVM). Moreover, as the minimum reference benchmark, we adopted the binary random model (RND) based on sorting Bernoulli binary random variables.

For models' performance comparison purposes, we adopted two metrics: (1) as the primary metric to assess the profit-generating potential of the predictions, the authors defined a metric (EARN), calculated according to Equation (1), to measure the earning of a theoretical USD/BRL trading strategy calculated using the predictions provided by the models; (2) to evaluate the predictions' correctness they adopted the Accuracy metric (ACC), the percentage of correct prediction.

$$EARN = \sum_{t=1}^{n} \widehat{Y}_t * v_t \qquad (1)$$

Where v_t is the USD/BRL variation on date t; $\hat{Y}_t$ is the prediction of the sign of v_t; and n is the number of trading days. By multiplying the model prediction of the sign of v_t ($\hat{Y}_t$) by the USD/BRL variation observed on the date t (v_t), we obtain the outcome of trading US$ 1.00 on the date (t-1) to earn R$ v_t on the date t. This value is positive if the prediction $\widehat{Y}_t$ is correct, and negative otherwise. To summarize the outcomes of the original work, we display the average EARN and ACC values produced by each model built in the original experiments in Table 1, henceforth named **Exp20.**

Table 1. EARN and ACC averages ordered by EARN (descending order) in Exp20

Model category	EARN	ACC
BMT	1.832	0.531
VOTING	1.477	0.516
RF	1.166	0.518
MLP	1.156	0.507
KNN	0.788	0.521
SVM	0.473	0.510
LOGREG	0.193	0.510
RND [a]	-0.081	0.495

[a.] Minimum benchmark.

In Table 2, we display the EARN average values generated by all the models and the improvement percentage provided by BMT over each of the other models' categories. On average, BMT generated EARN values 24% higher than VOTING, with 16% lower volatility. This outcome supported the central hypothesis of the research: the financial agents' collective actions affected by the trading calendar arguably induced deterministic patterns in the USD/BRL movement.

Table 2. EARN average (in descending order) achieved in Exp20 with the improvement % provided by BMT

Model category	EARN	The improvement % provided by BMT
BMT	1.832	
VOTING	1.477	24.0%
MLP	1.166	57.1%
RF	1.156	58.5%
KNN	0.788	132.5%
LOGREG	0.473	287.3%
SVM	0.193	849.2%
RND [b]	-0.081	2361.7%

[b] Minimum benchmark.

Nonetheless, these positive results raised the suspicion that, under this hypothesis, a naïve model (that takes the previous value as the prediction of the next) would also benefit from the calendar effect and potentially produce reasonable results, making it a challenging benchmark to consider. We were also concerned about the methodology's robustness during the pandemic. In the next section, we describe our methodology for investigating this guesswork.

3. METHODOLOGY

Motivated by the two questions stated at the end of the previous section, we considered verifying the robustness of the BMT during the pandemic and evaluating the naïve models (NAÏVE) performance. This session outlines the steps taken to make this verification. Nevertheless, to enhance clarity and ensure reproducibility, we first summarize the central concept of the method introduced in Section 2.

3.1 The Methodology Workflow

In our research, instead of retraining the models, we focus on enhancing performance by searching for temporal patterns in model behavior. Figure 1 shows the workflow diagram with the six main steps that comprise our method.

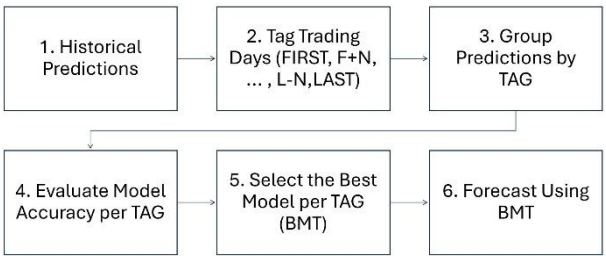

Figure 1. Methodology workflow

As noted in Section 2, our approach assumes the presence of the calendar effect. The only requirement for implementing our methodology is access to historical predictions from the trained voting-based ensemble model and all its constituent individual models (Step 1).

To facilitate a more effective comparison of model performance, we introduced a novel day-indexing framework and named it **day tagging**. Using the first and last trading days as references, we generate 22 unique tags, such as FIRST (first trading day), LAST (last trading day), and intermediate labels like F+1 through F+10 for days following the first and L-1 through L-10 for days preceding the last.

After tagging each trading day (Step 2), we use these labels to group the model's past predictions by day tag (Step 3). For each tag group, we evaluate all the individual models that constitute the ensemble and determine which model historically achieved

the highest prediction accuracy for that specific tag (Step 4). The process results in a tailored selection: based on past performance, we identify the most effective model for each tag (e.g., F+3 or L-2), the **Best Model per TAG**, **the BMT** (Step 5). This information is then used to generate predictions for future trading days with the same tag in the upcoming month (Step 6).

In the event of a tie between models, we default to the VOTING ensemble's prediction. Importantly, our approach is model-agnostic and can be applied to any ensemble configuration.

3.2 Description of the New Experiments

To assess the robustness of the BMT during the pandemic and evaluate the performance of NAÏVE models, we carried out new experiments, working with a more recent timeframe and adding the NAÏVE outcomes to the process. Henceforth, we name:

- **Exp22**: the **Exp20** experiments reproduced with more recent data.
- **NExp22**: the **Exp22** experiments, including the NAÏVE.

For **Exp22**, we repeated the two procedures detailed in subsection 3.1 by:

I. Building the VOTING using the following model categories: KNN, LOGREG, MLP, SVM, and RF.
II. Composing the BMT ensembles considering these 5 model categories plus the VOTING.

To verify the BMT robustness, we compared the outcomes in **Exp20** and **Exp22**, using ACC and EARN as performance metrics, and the RND's outcomes as the minimum benchmark.

In **NExp22**, the three updates described below were made to verify whether how much the NAÏVE impacts the process:

I. Including the NAÏVE predictions ($\hat{Y}_{t+1} = Y_t$) in the workflow.
II. Adopting the NAÏVE as the new minimum benchmark.
III. Changing the VOTING and to the BMT compositions' by adding the NAÌVE and removing the model category that produced the outcomes with the highest volatilities in Exp22**.**

Lastly, we compared the **Exp22** and **NExp22** performance metrics. The following section details the experiments highlighted above and presents the numerical results.

4. EXPERIMENTS DESCRIPTION

For comparison purposes, in the new set of experiments, **Exp22** and **NExp22**, we kept the same main components previously adopted in **Exp20:**

- The two performance metrics mentioned in Section 2 are the EARN metric, which measures the theoretical profit-generating potential provided by the models' outcomes, and the traditional Accuracy metric (ACC), which measures the models' overall performance.
- The five model categories listed in Section 2 (KNN, LOGREG, MLP, RF, and SVM) compose the VOTING.
- The RND category as the baseline benchmark.

4.1 Data Description and Preparation

We collected the data from the Institute for Applied Economic Research (IPEADATA)1, extracting 3,998 daily raw observations from April 4th, 2006, to March 31st, 2022. The upper plot in Figure 2 shows the daily USD/BRL (V_t), and the lower shows the curve of its variations ($v_t = V_t - V_{t-1}$).

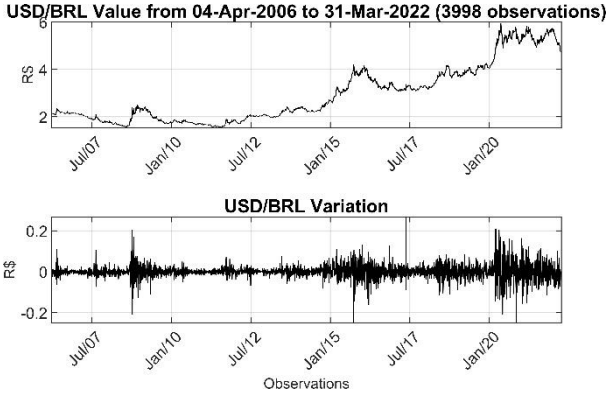

Figure 2. (a) The USD/BRL daily values (upper); (b) The USD/BRL daily variation (lower).

The basic statistics values of the cleaned-up USD/BRL daily variation time-series (v_t) are displayed in Table 3.

Table 3. USB/BRL variation (v_t) basic statistics

Basic Statistic	Value
Number of Observations	3.938
Mean	0.000
Standard Deviation	0.029
Minimum	-0.147
Maximum	0.155
Median	0.000
Volatility	61.775

We preprocessed the raw USD/BRL (V_t) variables by executing the following procedures:

1) Trading days: we considered only the trading days for the experiments, discarding weekends and holidays according to the Brazilian holiday calendar.
2) Outliers: as outlier identification criteria, we arbitrated a threshold value equal to 3.5 and discarded all the observations with USD/BRL log-variation values ($l_t = \log(V_t) - \log(V_{t-1})$) with Z-score values greater than it. In total, we discarded 59 observations.
3) Stationarity: we statistically tested and confirmed that the USD/BRL log-variation time-series (l_t) is stationary (in mean) using the Augmented Dickey-Fuller test with a 5% significance level and with the number of lagged difference terms equal to 252 (the number of trading days in one year).
4) Date Tagging: to better handle the difference in the trading days per month, we labeled the observations according to their relative position, using the first and last trading day of each month as references, using 22 tags defined as follows:

- FIRST: First trading day of the month.
- LAST: last trading day of the month.
- F+N: trading day occurring n-days after the F trading day of the month, with N assuming values from 1 to 10.
- L-N: trading day occurring n-days before the L trading day of the month, with N assuming values from 1 to 10.

4.2 Working Datasets

In this section, we describe how we composed the Working

[1] Retrieved on April 10th, 2023, from http://ipeadata.gov.br.

Datasets (WD), assuming all models have the same autoregressive formulation and use only time-series past values with no exogenous variables.

We defined the **output variable**, Y_t, as the sign of the USD/BRL variation, v_t, coding [-1] for negative variation and [+1] for positive variation. To compose the **input variables**, we arbitrated using the USD/BRL log-difference values (l_t), and worked with an observation window size equal to 10 (roughly half a month). Hence, each observation in the dataset ended up with the final structure:

- Output variable, Y_t: the sign v_t on date t.
- Input Variables, X_t: (l_{t-1}, ... l_{t-10}), ten autoregressive values of the l_t Log-variation values.
- Tag label, Tag_t: the TAG of the date t.
- Return value, Ret_t: the v_t value on date t.

We divided the complete dataset into 132 sliding (or rolling) windows, defining 132 WDs. Each WD is composed of a training dataset with 60 months (5 years) of observations and a test dataset with a month of observations.

We set the initial timeframe from May 16th, 2006, to March 31st, 2011. Shifting the previous time frame a month ahead defined the following sliding windows.

Table 4. Samples of the Training and the Test WD defined by sliding windows divisions (date and size)

WD ID	Training		
	Initial date	***Final date***	***# Obs.***
1	May 16th, 2006	Mar. 31st, 2011	1083
2	Jun. 8th, 2006	Apr. 29th, 2011	1098
…	…	…	…
131	Mar. 1st, 2021	Jan. 31st, 2022	1128
132	Apr. 3rd, 2017	Feb. 25th, 2022	1119
WD ID	**Test**		
	Initial date	Final date	# Obs.
1	Apr. 1st, 2011	Apr. 29th, 2011	19
2	May 2nd, 2011	May. 31st, 2011	22
…	…	…	…
131	Feb. 1st, 2022	Feb. 25th, 2022	14
132	Mar. 2nd, 2022	Mar. 31st, 2022	22

Then, we trained all the models using 60 months of observations to predict the observations of the following month. Table 4 displays a sample of the sliding window divisions and the size of the WDs. Due to space constraints, we listed only 4 of 132 Training and Test WDs.

4.3 Exp22: Reproducing Exp20 with more recent Data

For each of the 132 WDs defined as described in the previous section, we repeated the procedures specified in **Exp20** training five prediction model categories (KNN, LOGREG, MLP, RF, and SVM) according to the best practices acknowledged by the Machine Learning community[9].

These five model categories were used to compose the VOTING, using the majority voting criteria. To be referenced as the minimum benchmark, we built the RND using a Bernoulli binary random variables generator function to produce the predictions.

As noted in Section 2, we used the tag labels to separate the models' daily outcomes into 22 tag groups to obtain the BMT predictions. We then compared the ACC values generated by the models in the past per each of these 22 tags. We compared the five individual models' past ACC values and the VOTING to compose the BMT. In the event of a tie, we arbitrated the selection of the VOTING outcome. As the "past", we arbitrated the timeframe of the previous 12 WDs (previous year predictions), never including them in the model comparison process, preventing look-ahead bias problems.

Hence, the initial period corresponding to the first 12 WDs was not included in the overall evaluation. For this reason, we named 'Test timeframe' the period from the 13th to the 132nd, when we have the BMT predictions. Table 5 shows a summary of this division of timeframes.

Table 5. Timeframe division considered for performance evaluation

Timeframe	WD	Initial/Final Dates	# of Obs. (# years)
Total	1st ~ 132nd	Apr. 1st, 2011 / Mar. 31st, 2022	2623 (11 years)
"Past"	1st ~ 12th	Apr 1st, 2011 / Mar. 30th, 2012	251 (1 year)
Test	13th~ 132nd	Apr. 2nd, 2012 / Mar. 31st, 2022	2372 (10 years)

To compare the ACC values of the eight model categories (BMT, VOTING, KNN, LOGREG, MLP, RF, SVM, and RND), we concatenated the predictions obtained using the testing datasets from the Test Timeframe:

- 120 WDs (from the 13th to the 132nd WD).
- 2,372 aligned predictions, covering a period of 120 months (10 years) from April 2nd, 2012, to March 31st, 2022.

Table 6. Sliding window division sample

Sliding Window	Initial date	Final Date
#1	Apr. 2nd, 2012	Sep.18th, 2015
#2	Apr. 3rd, 2012	Sep. 21st, 2015
…	…	…
#1499	Oct. 16th, 2018	Mar. 30th, 2022
#1500	Oct.17th, 2018	Mar. 31st, 2022

To consistently compare the outcomes covering the whole Test Timeframe period, we calculated the ACC and the EARN values for 1,500 one-day sliding windows with 873 daily observations each. Each of the one-day sliding windows could be interpreted as the earnings path of a financial agent using the models' prediction to trade USD/BRL for 873 trading days (roughly equivalent to 41 months). Table 6 shows a sample of the sliding window division.

Table 7 displays the average EARN and ACC values produced by the eight model categories considering the 1,500 sliding windows. We listed the models sorted in descending order of the EARN average value.

Table 7. EARN and ACC averages ordered by EARN (descending order)

Model category	EARN	ACC
BMT	1.301 (G)	0.532 (G)
VOTING	1.203 (G)	0.523 (L)
RF	1.143 (G)	0.526 (G)
LOGREG	0.956 (G)	0.520 (G)
KNN	0.889 (E)	0.503 (L)

MLP	0.842 (G)	0.511 (L)
SVM	0.704 (G)	0.516 (G)
RND [c]	0.166	0.504

[c] Minimum benchmark.

We verified the model performance difference by applying the statistical test paired t-test with a 5% significance. We coded the test paired t-test outcomes besides the metric values as follows:

- (G) the model metric average is statistically greater than the next model on the list.
- (E) the model metric average is statistically equal to the next model on the list.
- (L) the model metric average is statistically lesser than the next model on the list.

Table 2 and Table 8, respectively, display the improvement provided by BMT in **Exp20** and **Exp22**. As expected, the lower values in Table 8 indicate a drop in the models' performance when the pandemic period data is included. Despite this, BMT still generated the highest metric value.

Table 8. EARN average (in descending order) achieved in Exp22 with the improvement % provided by BMT

Model category	EARN	The improvement % provided by BMT
BMT	1.301	
VOTING	1.203	8.2%
RF	1.143	13.8%
LOGREG	0.956	36.1%
KNN	0.889	46.3%
MLP	0.842	54.5%
SVM	0.704	84.9%
RND [d]	0.166	684.5%

[d] Minimum benchmark.

4.4 NExp22: Adding the NAÏVE to Exp22

Using the same 132 WDs described in Section 4.2, we defined the NAÏVE outcome as $\widehat{Y_t} = Y_{t-1}$, i.e., the prediction for the next USD/BRL variation signal is equal to the previous (or current) USD/BRL variation signal. Moreover, we adopted the NAÏVE outcomes as the new minimum benchmark.

As foreseen, on average, the EARN values generated by NAÏVE surpassed all the others, including the BMT's, as displayed in Table 9. Nevertheless, if we consider the model sequence ordered by ACC in descending order, NAÏVE would fall in the 5th position.

Table 9. EARN and ACC averages ordered by EARN (descending order)

Model category	EARN	ACC
NAÏVE [e]	1.968 (G)	0.517 (L)
BMT	1.301 (G)	0.532 (G)
VOTING	1.203 (G)	0.523 (L)
RF	1.143 (G)	0.526 (G)
LOGREG	0.956 (G)	0.520 (G)
MLP	0.889 (L)	0.503 (L)
KNN	0.842 (G)	0.511 (L)
SVM	0.704 (G)	0.516 (G)
RND	0.166	0.504

[e] Minimum benchmark.

Table 10 displays the improvement percentage provided by the NAÏVE outcomes over the other models. Also, on average, the NAÏVE EARN volatility was 18% lower than that of the BMT.

Table 10. EARN average (in descending order) achieved in NExp22 with the improvement % provided by NAÏVE

Model category	EARN	The improvement % provided by NAÏVE
NAÏVE [f]	1.968	
BMT	1.301	51.2%
VOTING	1.203	63.6%
RF	1.143	72.1%
LOGREG	0.956	105.8%
KNN	0.889	121.3%
MLP	0.842	133.6%
SVM	0.704	179.6%
RND	0.166	1086.4%

[f] Minimum benchmark.

This evidences that, on average, the NAÏVE predictions were correct for higher v_t values (USD/BRL variation) than the BMT, VOTING, RF, and LOGREG predictions. It also illustrates that only considering pure model performance metrics when dealing with trading applications may not be enough.

To observe the performance variation of the three main model categories (BMT, VOTING, and NAÏVE) over time, we plotted the curve of the ACC and the EARN values produced by the models in Figure 3 and Figure 4. Note that we are working with 1,500 one-day sliding windows, each representing the earning path of a financial agent using the models' prediction to trade USD/BRL for 873 days. In the graphs, we defined:

- The X-axis as the end date of the one-day sliding window.
- The Y-axis as the average value of the metric in the one-day sliding window.
- A horizontal dashed line to indicate minimum reference metric values:
- Y-axis equal to 0.5: below this line, the ACC average is below 0.5, indicating deficient performance.
- Y-axis equal to 0: below this line, the EARN average is negative, indicating trading loss.

For instance, the first point of the BMT curve in Figure 3 is (Sep.18th 2015, 0.5384). This information means that 0.5384 is the average ACC value achieved by BMT during the sliding window period ending on Sep.18th, 2015 (starting 873 working days earlier, on Apr. 2nd, 2012).

Applying the same principle, the first point of the BMT curve in Figure 4 is (Sep.18th, 2015, 0.7531). This information means that, on average, 0.7531 is the BRL(R$) amount earned by a financial agent daily trading USB/BRL using the BMT predictions during the sliding window period ending on Sep. 18th, 2015.

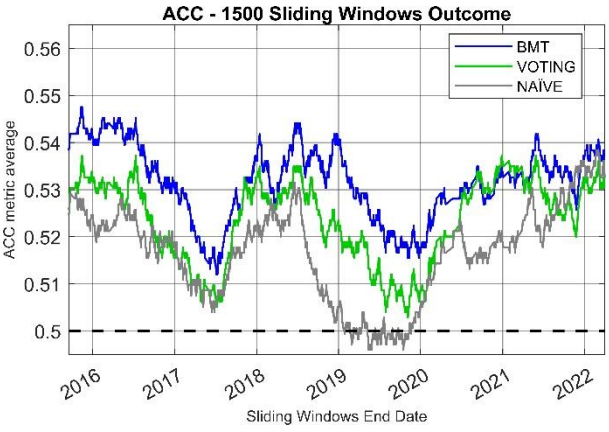

Figure 3. ACC average values produced by BMT, VOTING, and NAÏVE.

Figure 4 shows that the NAÏVE EARN curve always stays above the BMT's and VOTING's even though, in Figure 3, the NAÏVE ACC curve positioning is just the opposite, showing, again, how essential it is to have metrics to measure the potential earnings associated with methodologies for trading applications.

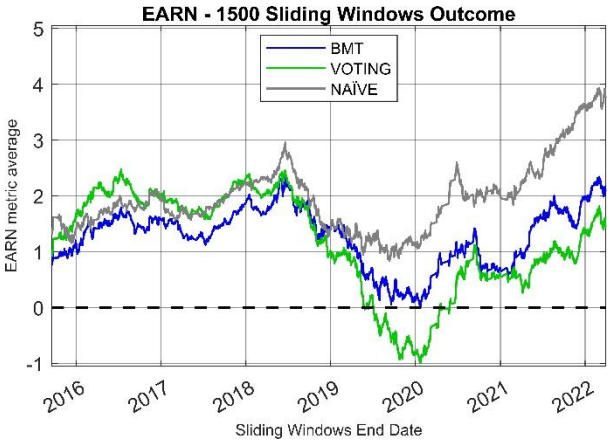

Figure 4. EARN average values produced by BMT, VOTING, and NAÏVE.

Since neither VOTING nor BMT have any restrictions regarding the individual models' composition, to take advantage of the good results produced by NAÏVE, we added the NAÏVE outcomes to the VOTING and to the BMT compositions, in both cases, replacing the MLP's. We chose MLP because, on average, its outcome generated the EARN values with the highest volatility.

Table 11 shows the EARN and ACC average values produced by the ensemble models (VOTING and BMT) in **NExp22**, including the NAÏVE outcomes in their compositions. Once again, as anticipated, replacing MLP with NAÏVE improved the performance of both the ensemble models.

Table 12 displays the percentage of improvement provided by BMT for the set of experiments performed with more recent data and the inclusion of NAÏVE in the ensemble model compositions. Comparing the **Exp20** outcomes with that of **NExp22**, the percentage of improvement provided by BMT over VOTING increased by 51.3% (from 24.0% to 36.3%, values from Table 2 and Table 12, respectively).

Table 11. EARN and ACC averages ordered by EARN (descending order) – with NAÏVE included in the BMT and the VOTING composition

Model category	EARN	ACC
BMT	2.710 (G)	0.541 (G)
VOTING	1.989 (E)	0.525 (G)
NAÏVE [g]	1.968 (G)	0.517 (L)
RF	1.143 (G)	0.526 (G)
LOGREG	0.956 (G)	0.520 (G)
KNN	0.889 (E)	0.503 (L)
MLP	0.842 (G)	0.511 (L)
SVM	0.704 (G)	0.516 (G)
RND	0.166	0.504

[g] Minimum benchmark.

Table 12. EARN average (in descending order) achieved in NExp22 with the improvement % provided by BMT – with NAÏVE included in the BMT and the VOTING composition

Model category	EARN	The improvement % provided by BMT
BMT	2.710	
VOTING	1.989	36.3%
NAÏVE [g]	1.968	37.7%
RF	1.143	137.1%
LOGREG	0.956	183.4%
KNN	0.889	204.7%
MLP	0.842	221.8%
SVM	0.704	285.1%
RND	0.166	1534.2

[g] Minimum benchmark.

To compare the **Exp22** and the **NExp22** outcomes, we displayed the EARN statistics generated by BMT and VOTING in the two sets of experiments in Table 13. The positive impact of including the NAÏVE outcomes can also be confirmed by observing the graphs in Figure 5 and Figure 6, which respectively show the ACC and the EARN curves produced by NAÏVE, VOTING, and BMT with the inclusion of the NAÏVE predictions in their compositions.

Table 13: EARN statistics generated by BMT and VOTING in Exp22 and NExp22, ordered by Average (in descending order)

Experiment / Model	EARN		
	Avrg.	Std.Dev.	Vol.*
NEXp2022 BMT	2.710	0.754	0.278
NEXp2022 VOTING	1.989	0.649	0.326
NEXp2022 NAÏVE	1.968	0.658	0.334
EXp2022 BMT	1.301	0.531	0.408
EXp2022 VOTING	1.203	0.899	0.747

*Vol. = (Std.Dev./Avrg.)

In Figure 5, the VOTING and the BMT curves went up away from the NAÏVE curve and the horizontal dashed line, the minimum model performance indicator (ACC average equal to 0.5). Likewise, in Figure 6Figure 6, for all the 1,500 one-day sliding windows end dates, the VOTING and the BMT curves ended significantly above the horizontal dashed line (EARN average equal to 0.0).

According to the outcomes obtained in **NExp22**, the predictions provided by the VOTING and BMT models, with the inclusion of the NAÏVE predictions, were able to consistently generate positive earnings for 10 years (the Test Timeframe period, detailed in Table 4), with BMT in the first place.

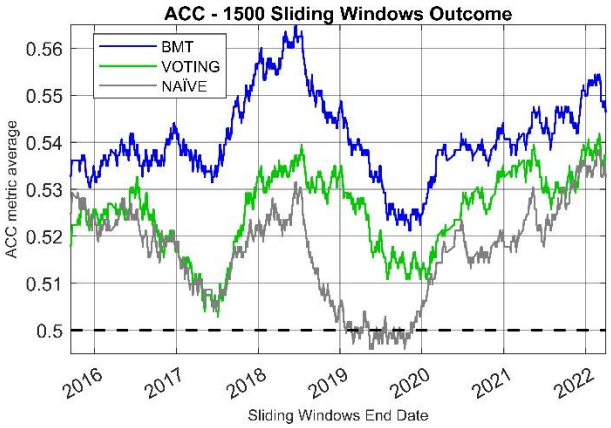

Figure 5. ACC average values produced by BMT, VOTING, and NAÏVE with the new configuration.

Figure 6. EARN average values produced by BMT, VOTING, and NAÏVE with the new configuration.

The code used to generate the results in this study was developed in MATLAB, and it is available in a Code Ocean capsule (link follows below). This capsule contains all necessary scripts and datasets required for reproducing the analysis. For any queries regarding the setup or usage of the provided code, please feel free to contact the corresponding author (https://codeocean.com/capsule/0312666/tree).

5. CONCLUSION

Our prior research served as the starting point for this present paper. In this earlier study, we investigated the benefits of adding the Behavioral Finance Theory perspective to the Machine Learning framework applied to the problem of predicting the next-day USD/BRL trend by proposing a method to improve existing voting-based ensemble models (VOTING) trained to provide this prediction.

The basic idea is that the calendar effects (a market anomaly acknowledged by the Behavioral Finance theory) affect investors' decisions, inducing deterministic patterns in the USD/BRL time-series and the outcomes of the models to forecast them. Therefore, by tracking these models' past performance, it would be possible to estimate their future performance and choose, among all available models, the 'best' one to provide the next prediction. Hence, this research proposed applying this strategy to improve a voting-based ensemble model and named it the Best Model per TAG ensemble model, BMT.

Despite the promising BMT results, we conjectured whether the calendar effect would also positively affect the predictions provided by naïve models (NAÏVE). We also anticipated that positive naïve models' outcomes could help to improve the outcomes generated by VOTING and BMT. Further, we wanted to verify the robustness of the BMT during the pandemic. For this paper, we thus collected more recent data, from April 2006 to March 2022, to include the pandemic and applied the BMT construction methodology by adding the NAÏVE's outcomes in the process.

We carried out a set of experiments and, as suspected, the NAÏVE's outcomes surpassed all the others, including the BMT's; we, therefore, included the former in the VOTING and the BMT compositions, and, as also expected, we got better results: comparing to the original experiments, the percentage of improvement provided by BMT over VOTING increased 51.3% (from 24.0% to 36.3%). Moreover, by comparing only the new experiments' outcomes, we got an average increase of 65.3% for VOTING and 108.3% for BMT.

Once again, our results supported the original research's initial hypothesis: assuming the existence of the calendar effect, it becomes possible to enhance the performance of trained voting-based ensemble models without retraining their individual components. Furthermore, we verified that the naïve model is not just a challenging benchmark; it is a model category that should be included in the BMT composition process to improve the outcomes of existing voting-ensemble models. We took this outcome as a reminder not to neglect simplicity.

We plan to continue our research by refining ensemble model construction methodologies and exploring the applicability of these methods to predict other financial assets in different financial markets.

6. ACKNOWLEDGMENTS

We thank Maria Cristina Vidal Borba for reviewing this paper. Our thanks also go to the ICONE-EPUSP Group, the NDS-EESP FGV Group, and Opencadd Advanced Technology for their support.

7. REFERENCES

[1] Abir, S.I. et al. 2024. Use of AI-Powered Precision in Machine Learning Models for Real-Time Currency Exchange Rate Forecasting in BRICS Economies. Journal of Economics, Finance and Accounting Studies. 6, 6 (2024), 66–83.

[2] Domingos, P. 1999. The Role of Occam's Razor in Knowledge Discovery. Data Mining and Knowledge Discovery. 3, 4 (1999), 409–425. DOI: https://doi.org/10.1023/A:1009868929893.

[3] Kailash Thiyagarajan 2025. Enhancing E-Commerce Product Page Recommendations using Large Language Models. International Journal of Computer Applications. 186, 68 (Feb. 2025), 49–54.

[4] López de Prado, M. 2018. Advances in financial machine learning. John Wiley & Sons.

[5] Matsumoto, E.Y. et al. 2022. Forecasting US Dollar Exchange Rate Movement with Computational Models and Human Behavior. Expert Systems with Applications. 194, (2022), 116521. DOI: https://doi.org/10.1016/j.eswa.2022.116521.

[6] Meese, R.A. and Rogoff, K. 1983. Empirical exchange rate models of the Seventies: do they fit out of sample? Journal of International Economics. 14, (1983), 3–24.

[7] Muhammad, A. et al. 2021. Forecasting Chinese Yuan/USD Via Combination Techniques During COVID-19. The Journal of Asian Finance, Economics and Business. 8, 5 (2021), 221–229. DOI: https://doi.org/10.13106/JAFEB.2021.VOL8.NO5.0221.

[8] Shobayo, O. et al. 2025. A Comparative Analysis of Machine Learning and Deep Learning Techniques for Accurate Market Price Forecasting. Analytics 2025, Vol. 4, Page 5. 4, 1 (Feb. 2025), 5. DOI: https://doi.org/10.3390/ANALYTICS4010005.

[9] Sicsú, A.L. et al. 2023. Técnicas de machine learning. Editora Blucher.

[10] Xie, Q. et al. 2024. FinBen: A Holistic Financial Benchmark for Large Language Models. Advances in Neural Information Processing Systems (2024), 95716–95743.

[11] Zhang, Y. and Hamori, S. 2020. The Predictability of the Exchange Rate When Combining Machine Learning and Fundamental Models. Journal of Risk and Financial Management. 13, 3 (2020), 48. DOI: https://doi.org/10.3390/JRFM13030048.