The effect of the COVID-19 pandemic to the WTI crude oil price using forecasting models

Marcelo Guarido, Daniel O. Trad, and Kristopher Innanen

ABSTRACT

Accurately forecasting the price of oil can be considered a Holy Grail in the petroleum industry, and even more important to take business decisions during a crisis, like the COVID-19 pandemic. For that, we are proposing the use of an ensemble of powerful forecasting methods to do an impact analysis of the pandemic to the oil and gas industry. By forecasting the prices with models trained with oil prices prior to 2020, we created a baseline of what the price of the WTI crude oil should have been without the pandemic and, with the information of the US production, we estimated that, in 6 months of the pandemic, a loss of around 60 billion USD in the US alone. We also estimate that, if the scenario of the pandemic does not change, the price of the oil for the 12 months after October 2020 will stay on a stable low level around 37 USD per barrel.

INTRODUCTION

Forecasting the oil price with high precision is ultimately important for financial decisions in the Oil&Gas industry, which is becoming more important every year, specially after the 2014 oil crisis (Norouzi et al., 2020), when the most recent one is due to the COVID-19 pandemic in 2020. Although, it is a difficult problem, as many different variables can impact the price of such volatile commodity (Bawks, 2020; Ma et al., 2019).

Several studies exists trying to make the best predictions possible, and most actually focus on the price history itself (used as a time series) to predict the future prices (Alquist et al., 2013). Gori et al. (2007) use linear and non-linear approximations curve fit to predict the yearly price and consumption of oil. Baumeister and Kilian (2015) test different approaches, combining external variables. Álvarez-Díaz (2020) uses parametric (such as ARMA, ARIMA, and FARIMA or ARFIMA) and non-parametric (nonlinear autoregressive neural network, genetic programming, and local regression K-NN) models, listing the advantages and disadvantages of them. Financial models, like the *gray economic model*, is used with relative success by Norouzi and Fani (2020). Long-memory (GARCH-M, or Generalized AutoRegressive Conditional Heteroskedasticity Model) and wavelet analysis are combined as a hybrid model by Lin et al. (2020). Machine learning models also play a role among several authors. Kulkarni and Haidar (2009) use neural networks as a non-linear forecasting model. Genetic algorithm-support vector machines (GA-SVM) is used as a regression model by Guo et al. (2012) to predict the price of oil for a week in advance. Zhang et al. (2015) decomposes the time series on several frequencies, or intrinsic mode functions (IMFs), plus a residual term, and combine the least square support machine to the GARCH model for a more robust forecasting. One thing in common of those papers is that none of them did prediction for the year 2020, when the price of the oil dropped drastically due to the COVID-19 pandemic, as most of the papers (even the ones publishes in 2020), were probably written before the latest crisis.

In this paper, we are forecasting the WTI crude oil price in 2020 first not considering the pandemic crisis and training the models, and them calculating the forecast for after the largest drop, estimating what was the loss in the Oil&Gas industry in the period from March to October of 2020. We are using monthly oil price data from $2005/01/03$ to $2020/10/19$, and compared the models *SARIMA*, *Facebook Prophet*, and *SARIMA + XGBoost*. In the last part, we will stack (ensemble) the 3 models into one for a more robust prediction.

THE DATA

The data was obtained from U.S. Energy Information Administration with help of an *API* key at the Quandl website. We download the daily price of WTI crude oil from $2005/01/03$ to $2020/10/19$ (Figure 1), with the barrel price is given in USD.

FIG. 1. Daily WTI crude oil price from 2005/01/03 to 2020/10/19.

The oil industry has been through several crisis in the past and in the present, the most recent one is the drop in prices due to the COVID-19 pandemic in 2020, with even an unlikely day when the price was negative ($2020/04/20$), with the barrel of oil costing -36.98 USD. As we are using models that only consider the past of the time series, without any external variable, we will remove, initially, the year 2020 from the data (Figure 2a), as this effect was never observed before for the oil price.

Another pre-processing was to smooth the time series by picking the average price of the month (Figure 2b), as the oil price can quite volatile, and a smooth version is more robust.

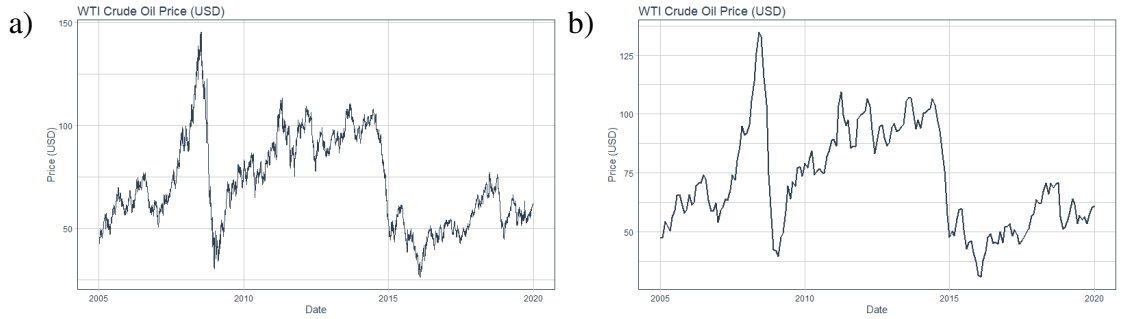

FIG. 2. a) Daily WTI crude oil price from 2005/01/03 to 2019/12/31 and b) the average monthly price for the same period.

FORECASTING MODELS

We are comparing different forecasting models: the *seasonal autoregressive integrated moving average*, or *SARIMA* (Cowpertwait and Metcalfe, 2009a), the *Facebook Prophet* model (Taylor and Letham, 2017), and a model that uses the *gradient boosting* from the *XGBoost* library (Chen and Guestrin, 2016) to boost the residuals of the SARIMA predictions.

Time Series Decomposition

For a given time series x_t, it can be decomposed into *additive* or *multiplicative* terms (Cowpertwait and Metcalfe, 2009b). The *additive decomposition* can be written as:

$$x_t = m_t + s_t + z_t \tag{1}$$

where t is the time, m_t is the trend, s_t is the seasonal term, and z_t is the error term (random, with normal distribution, and average equals to 0). The additive decomposition from equation 1 is used when the seasonal term do not increase with the trend. If the opposite occurs, then a *multiplicative decomposition* is more suitable (equation 2).

$$x_t = m_t \cdot s_t + z_t \tag{2}$$

In the case for the oil price, it was not observed an increase of seasonality amplitude when the trend increases, so the focus is to work with the additive decomposition of equation 1.

Figure 3 shows the decomposition of the oil price time series from Figure 1 (daily price). The plot on the top is the original time series, the second is the estimated *trend term*, the third is the *seasonal term* (set as 12 months), and the last plot shows the *errors term* (residuals). By using these decomposition that the *SARIMA* model will forecast the price of the oil.

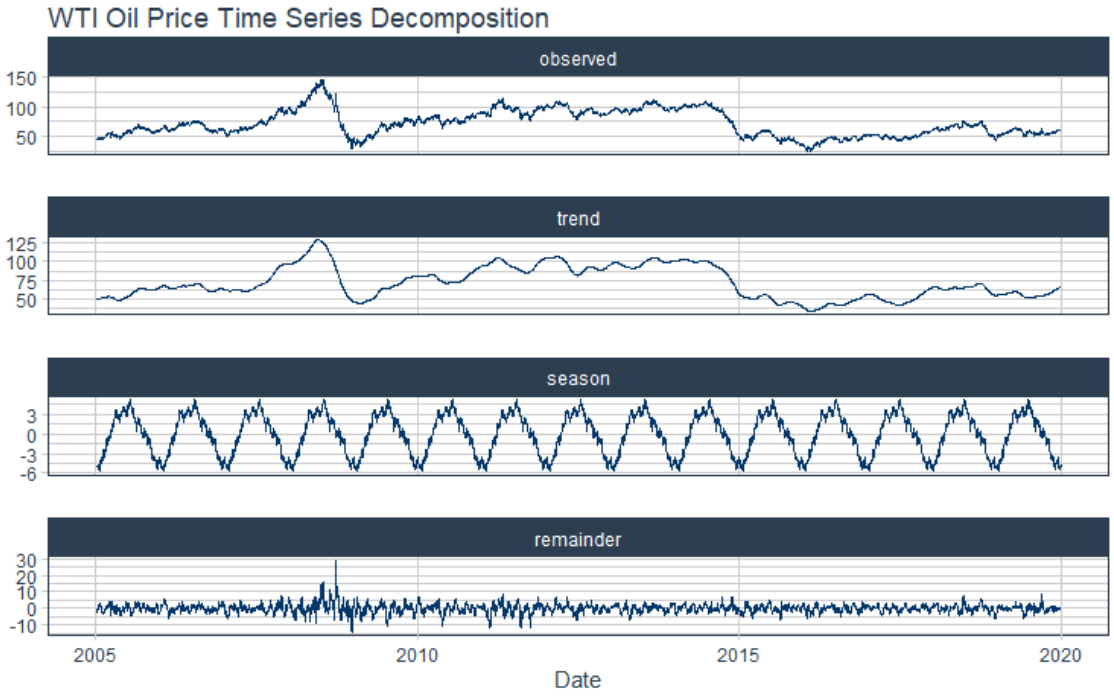

FIG. 3. Decomposition of the WTI crude oil price.

SARIMA (seasonal ARIMA)

SARIMA is the *autoregressive integrated moving average* (ARIMA) with a seasonal term Cowpertwait and Metcalfe (2009a). This model has several parameters for each part of the decomposition. The **autoregressive AR** part of order p of a time series x_t is a linear regression of the lags (past time $t-1, t-2, \ldots, t-p$) to a time t:

$$x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \cdots + \alpha_p x_{t-p} + w_t \tag{3}$$

where α_i are the models parameters and w_i is the white noise, and p can be interpreted as a window length of how much past data to be used for the regression. Equation 3 can be written as a polynomial of order p:

$$\theta_p(\mathbf{B})x_t = (1 - \alpha_1\mathbf{B} - \alpha_2\mathbf{B}^2 - \cdots - \alpha_p\mathbf{B}^p)x_t = w_t \tag{4}$$

where **B** is the backward shift operator, given by:

$$\mathbf{B}x_t = x_{t-1} \quad \text{and} \quad \mathbf{B}^n x_t = x_{t-n} \tag{5}$$

The **moving average MA** part of order q can be interpreted as a linear combination of the white noise (with zero mean and variance σ_w^2) with parameters β from the present to a past $t-q$:

$$x_t = w_t + \beta_1 w_{t-1} + \beta_2 w_{t-2} + \cdots + \beta_q w_{t-q} \tag{6}$$

Equation 6 can also be represented in terms of the operator **B**:

$$x_t = (1 - \beta_1 \mathbf{B} - \beta_2 \mathbf{B}^2 - \cdots - \beta_q \mathbf{B}^q) w_t = \phi_q(\mathbf{B}) w_t \tag{7}$$

An ARMA(p, q) model is given by:

$$\theta_p(\mathbf{B}) x_t = \phi_q(\mathbf{B}) w_t \tag{8}$$

The ARMA model is widely used for forecasting, but only works well on *stationary time series* (time series with no trend). The oil price time series contains a clear trend (Figure 3), so a transformation is required. That is when the **integrated I** part comes in hand. It computes the difference of a time series d times, until it is stationary (trend is removed):

$$(1 - \mathbf{B})^d x_t = w_t \tag{9}$$

Figure 4 show the conversion of a non-stationary time series to a stationary one.

The non-seasonal ARIMA(p, d, q) model is represented by:

$$\theta_p(\mathbf{B})(1 - \mathbf{B})^d x_t = \phi_q(\mathbf{B}) w_t \tag{10}$$

Equation 10 will perform a linear forecast of a time series. Now we want to include seasonal information, as seasonality is observed on the WTI crude oil price (Figure 3). The seasonal ARIMA (SARIMA) is an extension of the parameters (p, d, q) for the seasonal component. SARIMA includes autoregressive integrated moving average to the seasonal component with lag s (number of samples in one cycle), and has the notation SARIMA$(p, d, q)(P, D, Q)(s)$, and is expressed using the operator **B** as:

$$\Theta_P(\mathbf{B}^s)\theta_p(\mathbf{B})(1 - \mathbf{B}^s)^D(1 - \mathbf{B})^d x_t = \Phi_Q(\mathbf{B}^s)\phi_q(\mathbf{B}) w_t \tag{11}$$

where θ_P, θ_p, Φ_Q, and ϕ_q are, respectively, polynomials of order P, p, Q, and q.

Facebook Prophet

Facebook Prophet is forecast library for *R* and *Python* (Taylor and Letham, 2017), and is based on a generalized additive model (GAM) formulation (Hastie and Tibshirani, 1987). Given a time series x_t, it can be decomposed as:

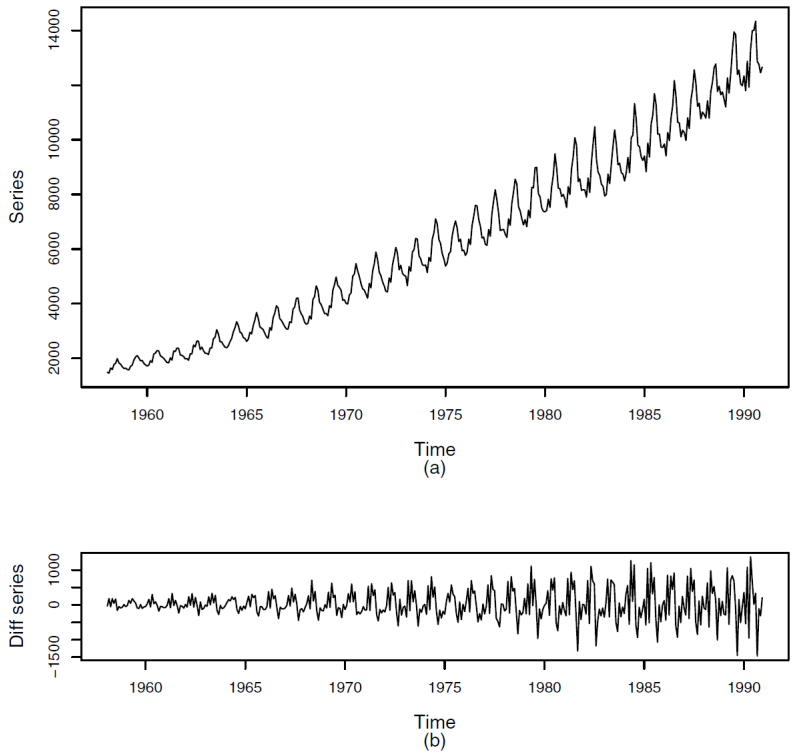

FIG. 4. Example of a) a time series with b) the trend removed (converted to a stationary time series) with the use of equation 9. Figure from Cowpertwait and Metcalfe (2009a).

$$x_t = m_t + s_t + h_t + z_t \tag{12}$$

where m_t is the trend, s_t is the seasonality (that can be split into hourly, daily, weekly, monthly, quarterly, and yearly seasonality in the library), h_t is the effects of holidays (non-regular events), and z_t is the error.

The trend is estimated with the help of *changepoints* (points where the trend can change) and is calculated by:

$$m_t = (k + \mathbf{a}_t^T \boldsymbol{\delta})_t + (o + \mathbf{a}_t^T \boldsymbol{\gamma})_t \tag{13}$$

where k denotes the base rate, $\boldsymbol{\delta}_t$ is a vector of rate adjustments, o is the offset parameter, $\boldsymbol{\gamma}_t$ is equals to $-c_t \boldsymbol{\delta}_t$ with c_t representing the changing point at a time t, and $\mathbf{a}_t$ is a vector that represents the continuity change of the base rate to its adjustment after a changing point. Between the changing points, the sub-trends are linear.

Taylor and Letham (2017) represent the seasonal term s_t as a standard Fourier series:

$$s_t = \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi n t}{P}\right) + b_n \sin\left(\frac{2\pi n t}{P}\right) \right) \tag{14}$$

where N is the number of observations in the period, P is the period (in days), and a and b are the fitting parameters. As an example, to construct a matrix for yearly seasonality and $N = 10$:

$$X_t = \left[\cos\left(\frac{2\pi(1)t}{365.25},\right), \ldots, \sin\left(\frac{2\pi(10)t}{365.25},\right)\right] \tag{15}$$

Then:

$$s_t = x_t\boldsymbol{\beta} \tag{16}$$

where $\boldsymbol{\beta} \sim \text{Normal}(0, \sigma^2)$ to smooth the seasonality term.

Holidays effects are added to the modeling by generating a matrix of regressors for each holiday i, assuming independence for each date, using D_i as a set of past and future dates for the holiday i, and assigning a parameter κ_i that corresponds to a change in the forecast due to the holiday. The matrix of regressors is then:

$$Z_t = [\mathbf{1}(t \in D_1), \ldots, \mathbf{1}(t \in D_L)] \tag{17}$$

The holiday term is then:

$$h_t = Z_t\boldsymbol{\kappa} \tag{18}$$

and as for the seasonality, a smooth parameter $\boldsymbol{\kappa} \sim \text{Normal}(0, \nu^2)$ is applied.

Gradient Boosting

Gradient Boosting is an ensemble tree method that can be used either for classification and regression (Chen and Guestrin, 2016). It is an iterative model on each the trees are generated one after the other in a way to optimize the previous tree. This means that each new tree is fitted on the residuals of the previous combination of trees. The model is well defined by Hastie et al. (2001).

In our forecast methodology, we use the gradient boost regressor as a support model by optimizing the residuals of the SARIMA model. This methodology is implemented in the library Modeltime for *R*.

Ensemble Models

The last model is an ensemble of the models list previously. For that we are using again the library Modeltime for *R*, as it offers ways to *stack* the models.

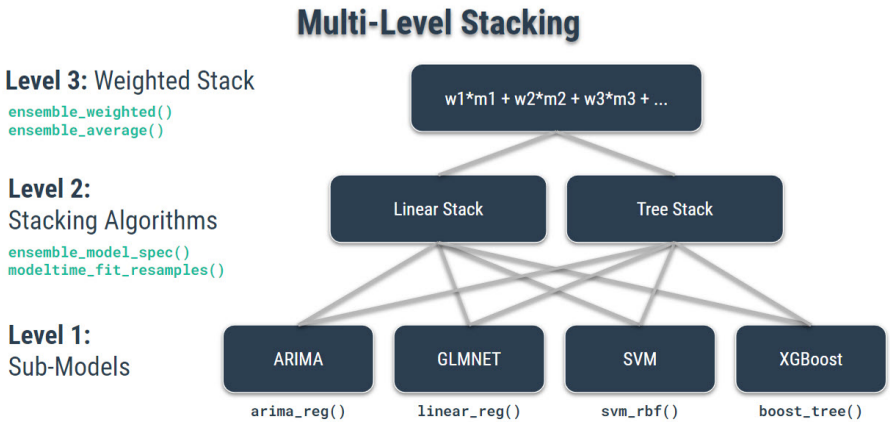

FIG. 5. Ensemble forecast diagram. In this project, we are using the *ensemble_average()* function. Figure modified from the Modeltime website.

Figure 5 is the schematic of the ensemble process. Different models which are, in our case, the *arima_reg()*, *arima_boost()*, and *prophet_reg()* are combined into 2 type of stacks: *linear* and *tree*. The last step is the average the stacks, and we used the function *ensemble_average()*.

The goal of averaging models is to retrieve the best of each one and create a precise and robust model in the end.

WTI CRUDE OIL PRICE FORECASTING ON THE COVID-19 PANDEMIC PERIOD

In this section, we present the forecasting with 4 different models: *Facebook Prophet*, *SARIMA*, *SARIMA + XGBoost*, and the *averaged model*, initially trained with data from $2005/01/01$ to $2019/12/31$, and see how the monthly price of oil would look like without the COVID-19 pandemic effect, estimating what is the loss against the true prices. Later, we retrain all models with data from $2005/01/03$ to $2020/10/19$ (with the pandemic period included), and forecast the monthly price of oil 12 month in the future.

Forecasting 2020's Oil Price without the Pandemic Effect

To evaluate the impact of the COVID-19 pandemic on the Oil&Gas industry, first we removed from the training set any price related to the year 2020. All the models are then trained with the remaining data, ending up by forecasting the price for the 10 first month of 2020.

Figure 6 shows the predictions of the 4 trained models. In both *SARIMA* models, the algorithm automatically determine the model parameter (p, d, q, P, D, Q), but the seasonality is an input parameter, which, by trail and error, was set to 6 months. In both cases, for the non-seasonal part, the found parameters were $p = 1$, $d = 1$, and $q = 0$, meaning that the model required a polynomial of order 1 for the auto-regressive ($p = 1$), required 1 differentiation to make the series stationary ($d = 1$), and did not need a moving-average factor ($q = 0$). For the seasonal part, only an autoregressive term was required ($P = 1$), suggesting the presence of seasonality in the data (which we observed on Figure 3). The

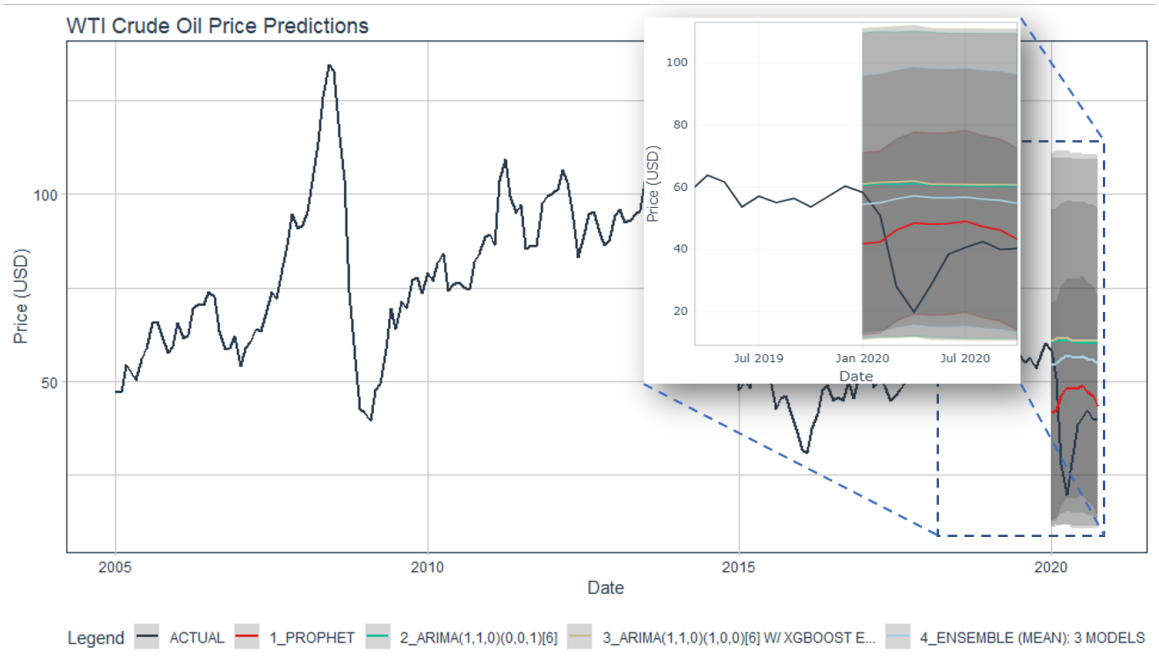

FIG. 6. Monthly WTI crude oil price from $2005/01$ to $2020/10$ compared with the forecasting models.

difference from the SARIMA models is that one of the *XGBoost* model fitted on the residuals. Both models showed very similar behavior, with the SARIMA + XGBoost having slightly more seasonality. And both models are predicting values higher than the other models, showing an optimistic scenario. The *Facebook Prophet* model (red in Figure 6) was trained considering monthly and yearly seasonality, while the trend is estimated with use of 25 change points. Predictions for this model show more variation, which is promising for the data, but with very pessimistic prices, as the model was predicting the crisis, which is improbable. Note the high drop in values on the first prediction point (January) to the last true price (December). Averaging the models (light blue in Figure 6) created a more realist model (when considering continuity with the previous months data), with a fairly optimistic model (less than the SARIMAs), but also including some of the variation from the Prophet, looking a more robust model. Evaluation of the monetary loss in the Oil&Gas is done using the ensemble model.

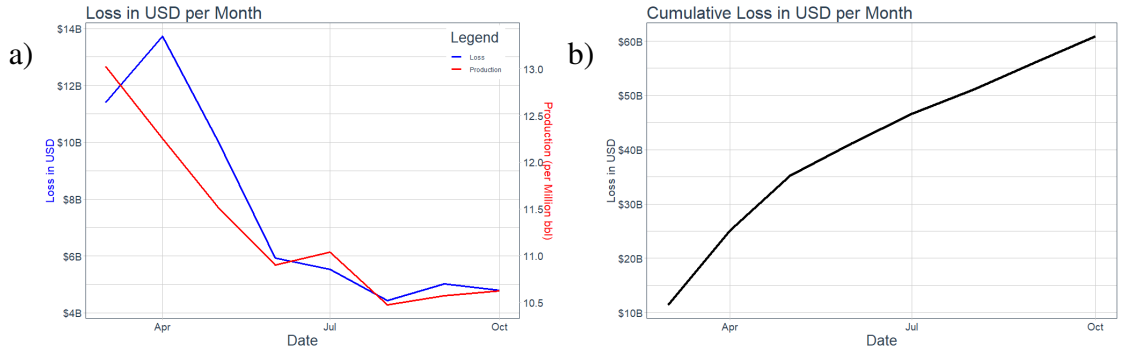

FIG. 7. a) Monthly loss (blue) and production (red) and b) cumulative monthly loss.

Figure 7 shows the a) monthly loss (blue) and production (red) and b) cumulative

monthly loss. The US weekly production data (which is the average of daily production for each week) was downloaded from the U.S. Energy Information Administration website. To match the price and production data, we averaged the weekly (daily averaged) production to monthly average. In countries like Canada and USA, the COVID-19 spreading was officially declared as pandemic in March 2020, so we are estimating the loss from March to October 2020. In Figure 7a, we see a high drop in production after March, with a lower rate of drop trend during the subsequent months. In the same plot, there is the loss curve, which represents the total loss in dollars due to the pandemic for each month. To create this curve, the average production, difference between true and predicted prices, and the number of days for each month are multiplied, generating an brute loss for the month. The is higher in the first couple month of the pandemic, and the monthly loss decrease as the production decreases. In the Figure 7b, is the cumulative loss in USD after each month. With our methods for forecasting and loss calculation, we are estimating a loss of around USD 60 billion until the end of October in the US alone.

Forecasting with Pandemic Effect

Now using the price of WTI crude oil during the pandemic, we train our models for the future 12 months from November 2020 (Figure 8). In this scenario, all the models are in the same level (any model is more optimistic or pessimistic), starting the forecast with approximately the same price, but with the *Facebook Prophet* still presenting more variability, ending the forecasting with a descending trending. The *SARIMA* and *SARIMA + XGBoost* models are still close to each other, with low variability. The *ensemble* model is closer to the SARIMAs ones, but capturing some of the variability from the Facebook Prophet.

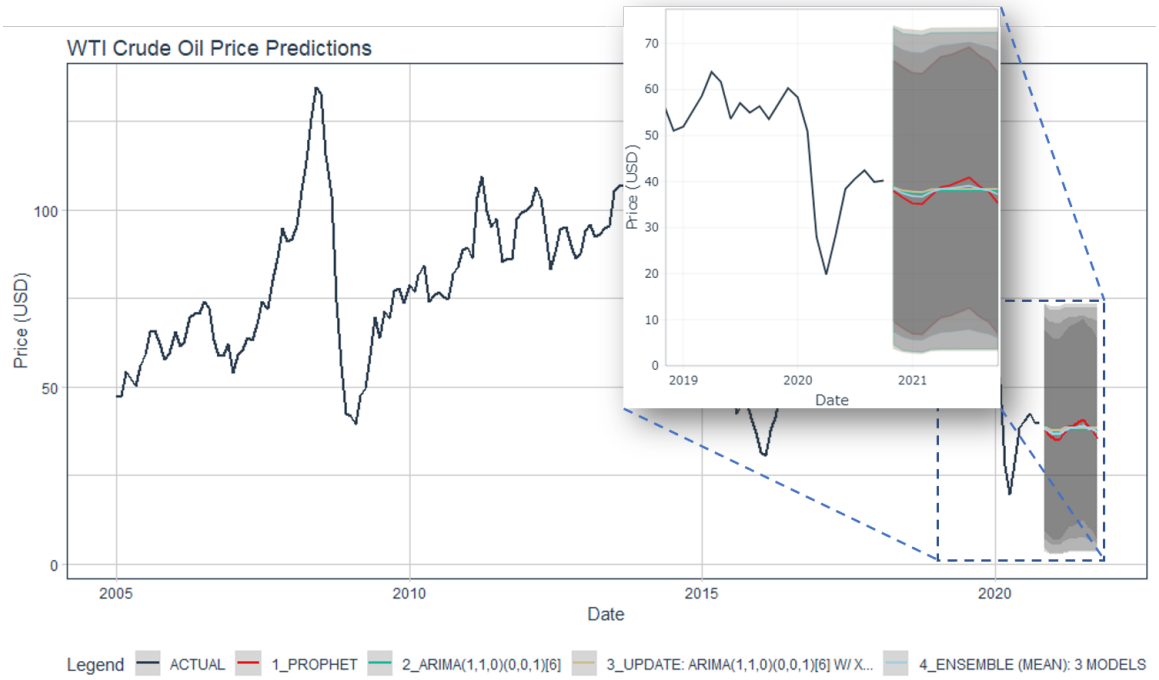

FIG. 8. Monthly WTI crude oil price from $2005/01$ to $2020/10$ (in black) and one year forecasting with the 4 trained models.

If the pandemic scenario continues stable, our forecasting presents an also stable scenario for the WTI crude oil prices, but in a low level if compared to previous year, around US$37 per barrel.

CONCLUSIONS

In this report, we presented different methods to forecast time series data, and we applied them to the WTI crude oil price. all the presented models showed advantages and disadvantages, but by averaging them, we ended up with a robust ensemble model, which we used for our impact analysis.

By comparing the predicted prices of the crude oil against the true prices in 2020, and including the US production data, we did the impact analysis of the COVID-19 pandemic to the Oil&Gas industry and estimated a loss of around US$60 billions from March to October of 2020 (a period of 6 months) in the US alone.

In the end, we forecast the price of oil from November 2020 to 12 months in the future and conclude that, if the pandemic scenario does not change, the price of the oil will continue stable on a low level, around US$37 per barrel.

ACKNOWLEDGEMENTS

The authors thank the sponsors of CREWES for continued support. This work was funded by CREWES industrial sponsors and NSERC (Natural Science and Engineering Research Council of Canada) through the grant CRDPJ 461179-13 and CRDPJ 543578-19, and the financial support from Canada First Research Excellence Fund (CFREF).

REFERENCES

Alquist, R., Kilian, L., and Vigfusson, R. J., 2013, Chapter 8 - forecasting the price of oil, *in* Elliott, G., and Timmermann, A., Eds., Handbook of Economic Forecasting, Elsevier: Handbook of Economic Forecasting, **2**, 427 – 507.

Álvarez-Díaz, M., 2020, Is it possible to accurately forecast the evolution of brent crude oil prices? an answer based on parametric and nonparametric forecasting methods: Empirical Economics, **59**, No. 3, 1285–1305.

Baumeister, C., and Kilian, L., 2015, Forecasting the real price of oil in a changing world: A forecast combination approach: Journal of Business & Economic Statistics, **33**, No. 3, 338–351.

Bawks, B., 2020, U.s. energy information administration - eia - independent statistics and analysis.
URL `https://www.eia.gov/finance/markets/crudeoil/`

Chen, T., and Guestrin, C., 2016, Xgboost: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
URL `http://dx.doi.org/10.1145/2939672.2939785`

Cowpertwait, P. S., and Metcalfe, A. V., 2009a, Non-stationary models, *in* Introductory Time Series with R, chap. 7, Springer, 137–157.

Cowpertwait, P. S., and Metcalfe, A. V., 2009b, Time series data, *in* Introductory Time Series with R, chap. 1, Springer, 1–25.

Gori, F., Ludovisi, D., and Cerritelli, P., 2007, Forecast of oil price and consumption in the short term under three scenarios: Parabolic, linear and chaotic behaviour: Energy, **32**, No. 7, 1291 – 1296.

Guo, X., Li, D., and Zhang, A., 2012, Improved support vector machine oil price forecast model based on genetic algorithm optimization parameters: AASRI Procedia, **1**, 525 – 530, aASRI Conference on Computational Intelligence and Bioinformatics.

Hastie, T., and Tibshirani, R., 1987, Generalized additive models: Some applications: Journal of the American Statistical Association, **82**, No. 398, 371–386.

Hastie, T., Tibshirani, R., and Friedman, J., 2001, The elements of statistical learning - data mining, inference, and prediction: Springer, second edn.

Kulkarni, S., and Haidar, I., 2009, Forecasting model for crude oil price using artificial neural networks and commodity futures prices, 0906.4838.

Lin, L., Jiang, Y., Xiao, H., and Zhou, Z., 2020, Crude oil price forecasting based on a novel hybrid long memory garch-m and wavelet analysis model: Physica A: Statistical Mechanics and its Applications, **543**, 123,532.

Ma, Y.-r., Ji, Q., and Pan, J., 2019, Oil financialization and volatility forecast: Evidence from multidimensional predictors: Journal of Forecasting, **38**, No. 6, 564–581.

Norouzi, N., and Fani, M., 2020, Black gold falls, black plague arise - an opec crude oil price forecast using a gray prediction model: Upstream Oil and Gas Technology, **5**, 100,015.

Norouzi, N., Fani, M., and Ziarani, Z. K., 2020, The fall of oil age:a scenario planning approach over the last peak oil of human history by 2040: Journal of Petroleum Science and Engineering, **188**, 106,827.

Taylor, S. J., and Letham, B., 2017, Forecasting at scale: PeerJ Preprints, **5**, e3190v2.

Zhang, J.-L., Zhang, Y.-J., and Zhang, L., 2015, A novel hybrid method for crude oil price forecasting: Energy Economics, **49**, 649 – 659.