Customs revenue prediction using ensemble methods (statistical modeling vs machine learning)

Jordan Simonov and Zoran Gligorov

PICARD 2020

23-26 November 2020

Introduction

Forecasting problem

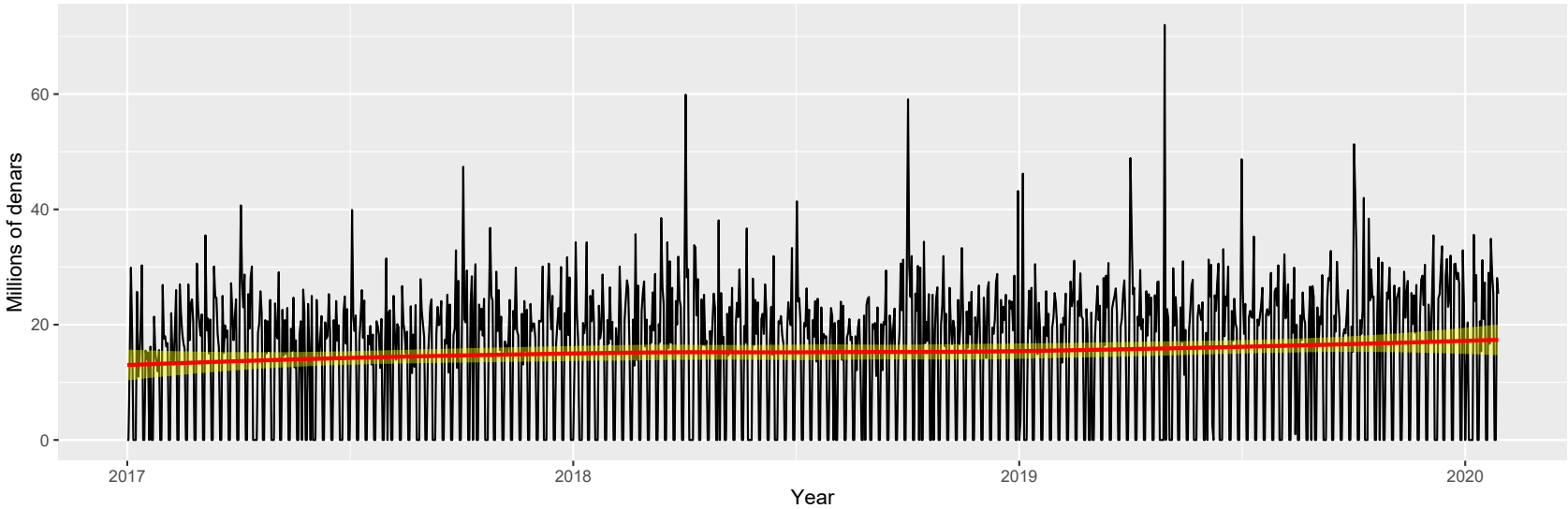
- Finance ministries usually project the annual and monthly revenue collection targets that are expected to be met by customs administrations
- Modern cash management requires accurate short-term forecasts not only on a monthly basis, but also on a weekly or daily basis
- Forecasting daily collection of customs revenues
- Possible approaches?

Table: Statistical modelling vs Machine learning

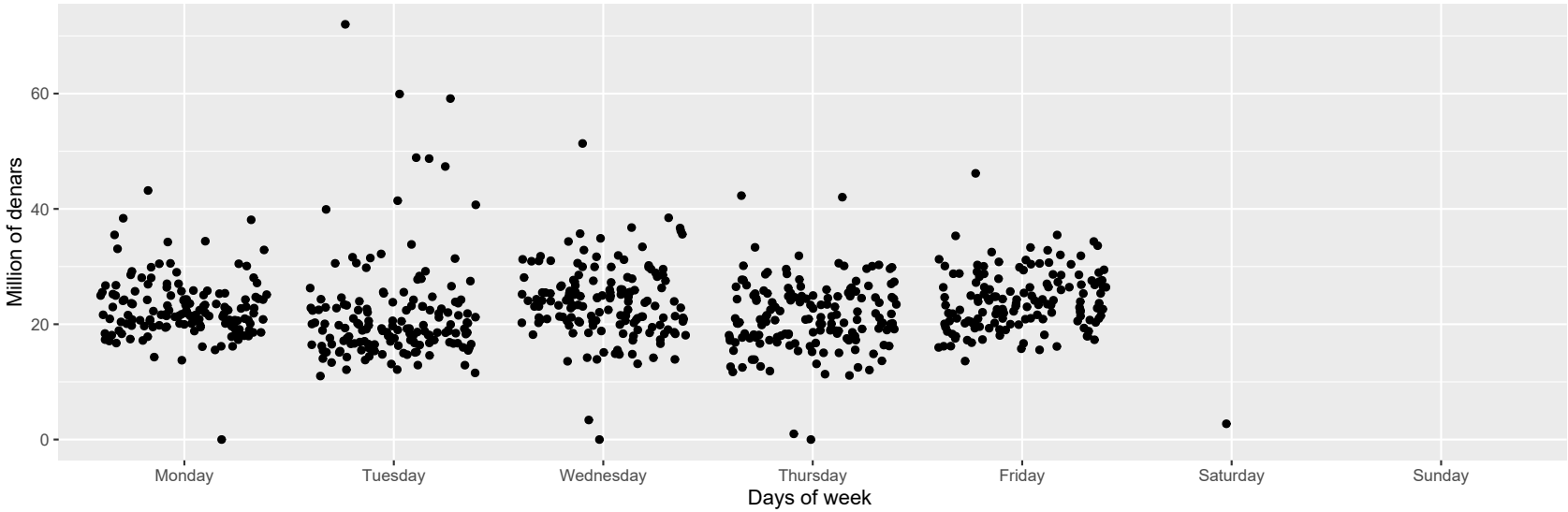
Statistical modelling	Machine learning
Formalisation of relationships between variables in the form of mathematical equations.	Algorithm that can learn from the data without rule-based programming.
Required to assume shape of the model curve prior to performing model fitting to the data (e.g. linear, polynomial).	Does not need to assume underlying shape, as machine learning algorithms can learn complex patterns automatically, based on the provided data.
Predicts the output with 85% accuracy at a 90% confidence level.	Predicts the output with 85% accuracy.
Various diagnostics of parameters are performed, such as p-value.	Does not perform statistical diagnostic significance tests.
Data will be split into 70%/30% to create training and testing data. Model developed on training data and tested on testing data.	Data will be split into 50%, 25%/25% to create training, validation, and testing data. Models developed on training and hyperparameters are tuned on validation data and are evaluated against test data.
Models can be developed on a single dataset (training data), as diagnostics are performed at both overall accuracy and individual variable level.	Need to be trained on two datasets (training and validation data), to ensure two-point validation.
Mostly used for research purposes.	Apt for implementation in a production environment.
From the school of statistics and mathematics.	From the school of computer science.

Source: Adopted according to Pratap (2017, p. 43)

Daily customs duties collection in period 2017-2020, in MKD denars



Customs duties daily collection in period 2017–2020, in MKD denars



Statistical modelling

Statistical modelling

Statistical modelling

- **Exponential smoothing (ETS)**

Statistical modelling

- **Exponential smoothing (ETS)**
- **Auto-regressive integrated moving average (ARIMA)**

Statistical modelling

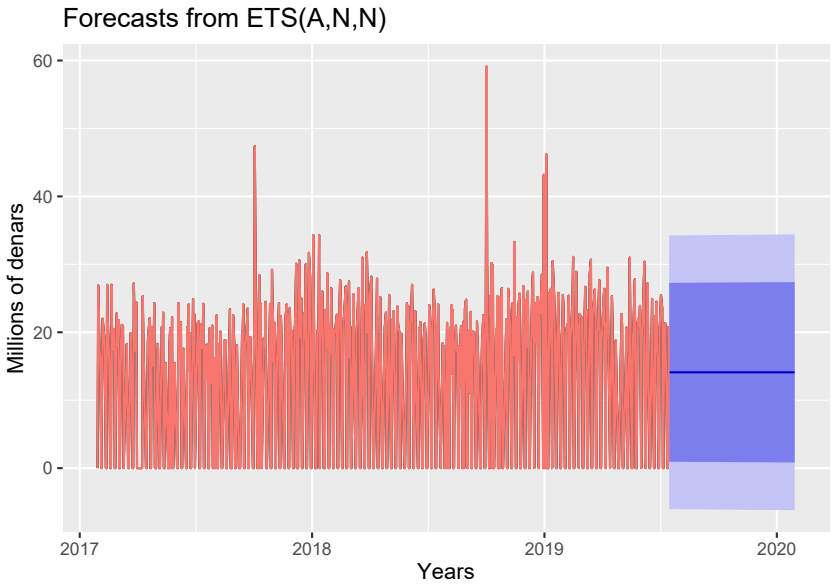
- **Exponential smoothing (ETS)**
- **Auto-regressive integrated moving average (ARIMA)**
- **Forecasting with decomposition (STL)**

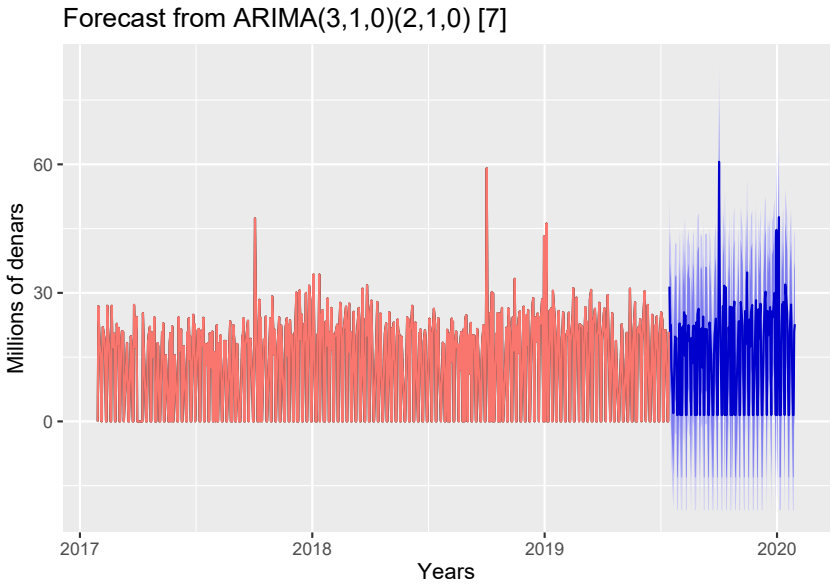
Statistical modelling

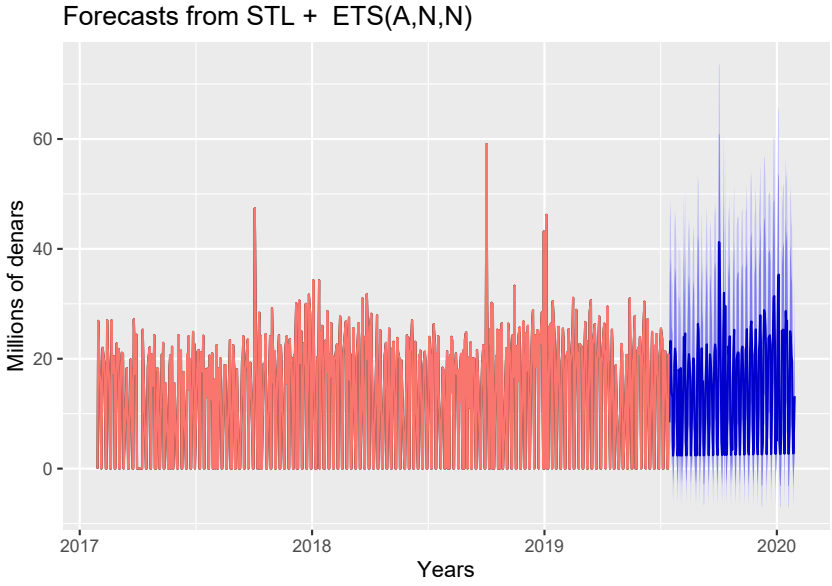
- **Exponential smoothing (ETS)**
- **Auto-regressive integrated moving average (ARIMA)**
- **Forecasting with decomposition (STL)**
- **Trigonometric Exponential smoothing state space model with Box-Cox transformation, ARMA errors (TBATS)**

Statistical modelling

- **Exponential smoothing (ETS)**
- **Auto-regressive integrated moving average (ARIMA)**
- **Forecasting with decomposition (STL)**
- **Trigonometric Exponential smoothing state space model with Box-Cox transformation, ARMA errors (TBATS)**


Forecasts from ETS(A,N,N)
Millions of denars
Years
2017
2018
2019
2020
0
20
40
60


Forecast from ARIMA(3,1,0)(2,1,0) [7]
Millions of denars
Years
2017
2018
2019
2020
0
30
60


Forecasts from STL + ETS(A,N,N)
Millions of denars
Years
2017
2018
2019
2020
0
20
40
60

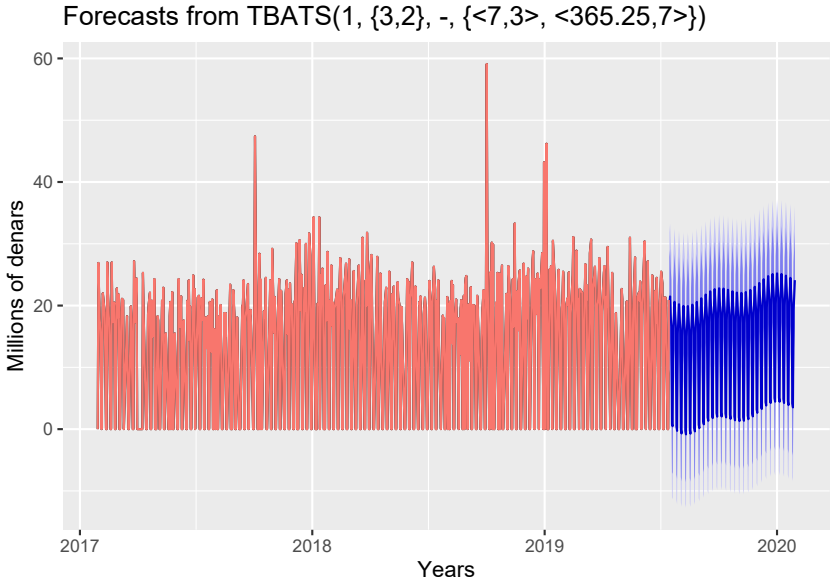

Forecasts from TBATS(1, {3,2}, -, {<7,3>, <365.25,7>})
Millions of denars
Years
2017
2018
2019
2020
0
20
40
60

Table: Evaluating forecast accuracy

	Training set		Test set	
	RMSE	**MAE**	**RMSE**	**MAE**
ETS	11.71	10.02	12.89	11.79
STL	3.04	2.33	6.08	4.70
ARIMA	8.7	5.96	11.46	8.57
TBATS	4.64	3.39	5.23	4.00

Machine learning

Machine learning

- **Linear regression (LM)**

Machine learning

- **Linear regression (LM)**
- **Classification and regression trees (CART)**

Machine learning

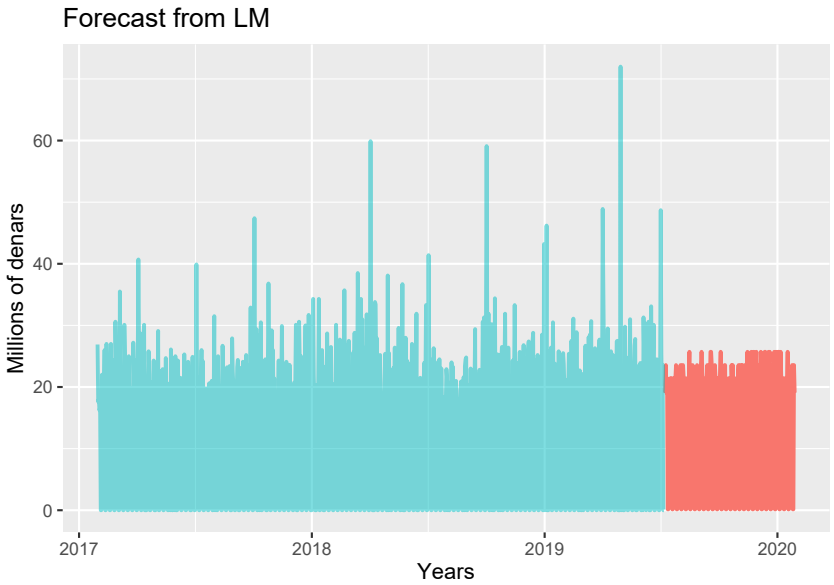
- **Linear regression (LM)**
- **Classification and regression trees (CART)**
- **Conditional inference tree (CTREE)**

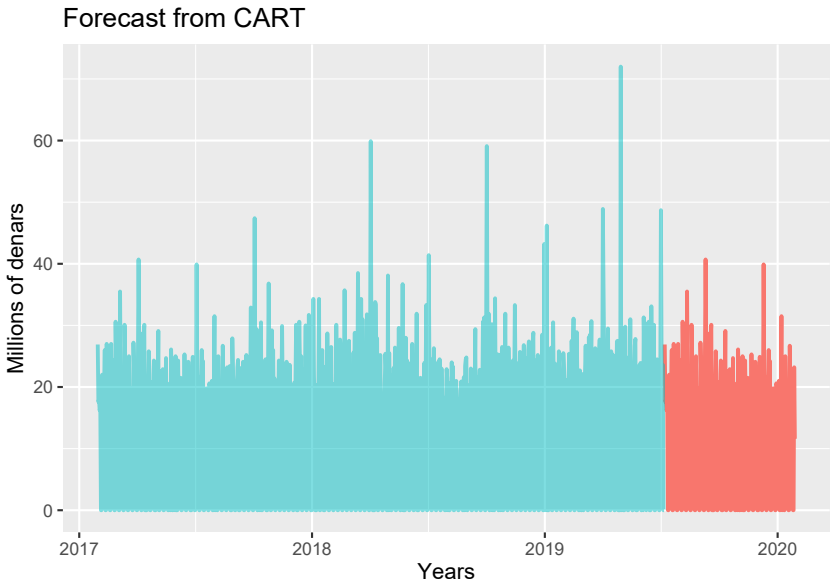
Machine learning

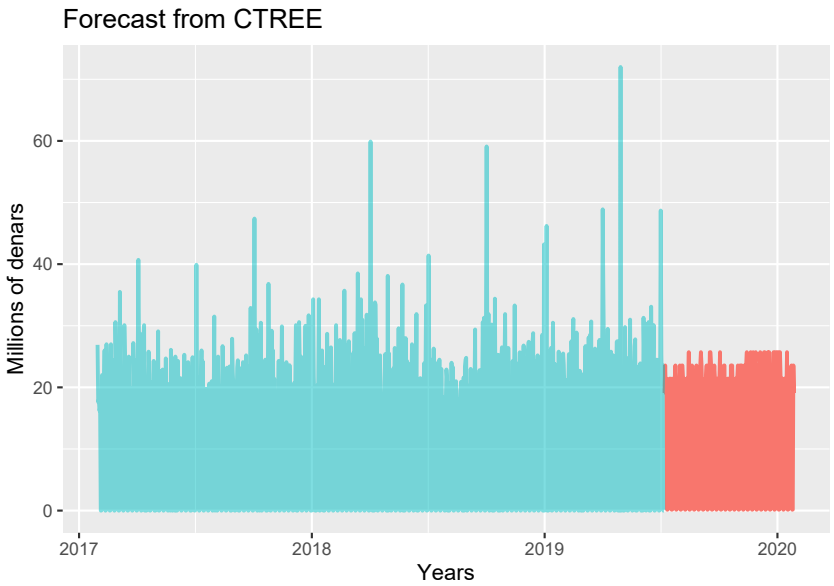
- **Linear regression (LM)**
- **Classification and regression trees (CART)**
- **Conditional inference tree (CTREE)**
- **eXtreme Gradient Boosting (XGBoost)**

Machine learning

- **Linear regression (LM)**
- **Classification and regression trees (CART)**
- **Conditional inference tree (CTREE)**
- **eXtreme Gradient Boosting (XGBoost)**


Forecast from LM
Millions of denars
Years
0
20
40
60
2017
2018
2019
2020


Forecast from CART
Millions of denars
Years
0
20
40
60
2017
2018
2019
2020


Forecast from CTREE
Millions of denars
Years
0
20
40
60
2017
2018
2019
2020

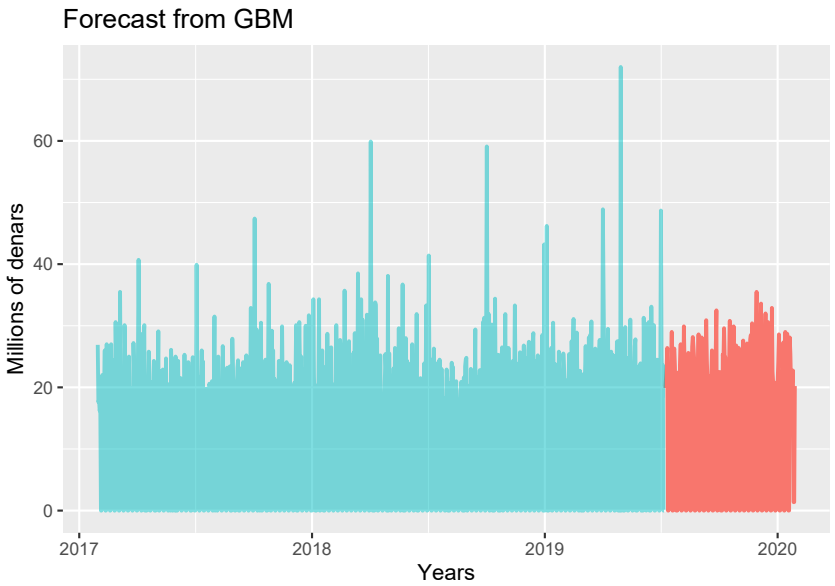
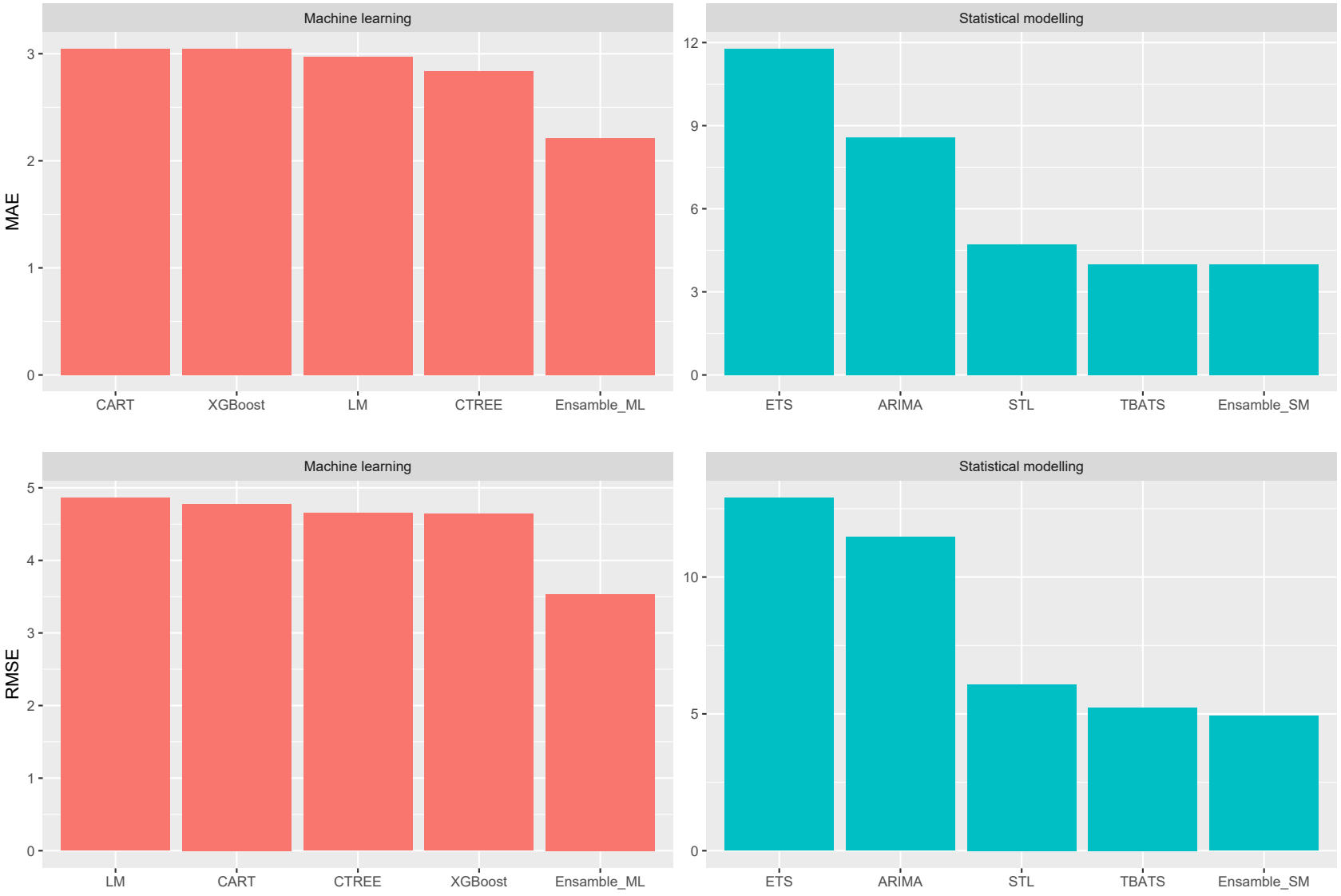

Forecast from GBM
Millions of denars
Years
0
20
40
60
2017
2018
2019
2020

Table: Evaluating forecast accuracy

	Training set		Test set	
	RMSE	MAE	RMSE	MAE
LM	3.81	2.54	4.86	2.97
CART	4.27	2.72	4.78	3.05
CTREE	4.15	2.52	4.65	2.84
XGBoost	3.80	2.41	4.64	3.05

Ensamble learning

Evaluating forecast accuracy



- **By using statistical modelling and machine learning, we tested 10 different models in the R ecosystem**

- **By using statistical modelling and machine learning, we tested 10 different models in the R ecosystem**
- **Using statistical modelling, the ensemble technique reduced the RMSE error from 12.99 to 4.92, while when using machine learning, the error of 4.86 was reduced to 3.53**

Conclusion

Conclusion

- As for both approaches, the ensemble technique has shown that it can improve prediction accuracy compared to the individual models. They can reduce the forecasting error, so to that end, we can conclude that the ensemble technique is certainly a game changer and must be an important addition to every forecaster's toolbox. For this reason, we recommend using this technique for forecasting purposes with statistical modelling or machine learning. The choice is yours!

References

References

- Pratap, D. (2017). Statistics for machine learning. Packt Publishing Ltd.
- Nielsen, A. (2019). Practical time series analysis: Prediction with statistics and machine learning. O'Reilly Media, USA
- Hyndman. R. J., & Athanasopoulos, G. (2016). Forecasting: Principles and practice. Monash University, Australia
- Lewis,N.D (2017). Neural Networks for Time-Series Forecasting with R