GROUP WORK PROJECT # 3
GROUP NUMBER: 6442

MScFE 660: RISK MANAGEMENT

FULL LEGAL NAME	LOCATION (COUNTRY)	EMAIL ADDRESS	MARK X FOR ANY NON-CONTRIBUTING MEMBER
Marin Stoyanov	Bulgaria	azonealerts@gmx.com	
Prabhdeep Kaur	India	prabhdeep089kaur@gmail.com	

Statement of integrity: By typing the names of all group members in the text boxes below, you confirm that the assignment submitted is original work produced by the group (excluding any non-contributing members identified with an "X" above).

Team member 1	Marin Stoyanov
Team member 2	Prabhdeep Kaur
Team member 3	

Use the box below to explain any attempts to reach out to a non-contributing member. Type (N/A) if all members contributed.
Note: You may be required to provide proof of your outreach to non-contributing members upon request.

N/A

Introduction

In this mini-capstone project, we are working on the problem of underlying factors driving crude oil prices. For this, we have used the macroeconomic and financial datasets and we'll proceed with the same data in this section as well. In continuation with GWP1 and GWP2, this section will provide the improvement of the model and interpretation of the results.

Results And Interpretation

Purpose of training dataset

The training set is important because this is where the model will learn the dependencies between the data values and then it will use this knowledge to make the statistical inference and apply it into unseen data.

- In the case of supervised learning this data will be labeled and here the model will find the connections between the data points, create, refine rules and fit its parameters towards the things that it has learned from the dataset.
- In the opposite case (unsupervised learning) the model will use unlabeled data and it will find patterns in the data or also called "behavior".
- In both cases training data is essential for the learning process because it is the foundation and the place where to learn the proper model behavior which will be applied towards the unseen data where the predictions will be made [1]

Purpose of validation dataset

The validation dataset is an important component for the development of the models whether econometric, financial modeling or other data-driven models. Validation dataset serves as an intermediary between the training and the testing dataset. It ensures the generalization of the model to the new dataset.

Validation dataset plays the following important roles-

- Validation dataset are useful especially when tuning the hyperparameters. This is where we are trying to improve the model's accuracy and avoid overfitting or underfitting the training dataset. These can be done by adjusting the hyperparameters when evaluating the model's performance over the validation set. This is the main purpose of this set.

- Validation dataset also plays a vital role in preventing overfitting. It is used as a gauge indicator for the model's performance over unseen data and this is where we can detect and see if it overfits or not. The training should be stopped once the performance starts degrading to avoid overfitting.
- Monitoring the performance of the model is very important because it gives an indication that maybe there is a problem with the data and this is why the model is underperforming. In such cases additional adjustments may be done (based on the metrics derived from testing the model over the validation dataset) in order to enhance the performance.

Validation dataset is crucial to ensure that the performance of the model is optimized before making final evaluation on the testing dataset. It leads to better predictive performance by doing hyperparameter tuning and validation on a separate dataset. [1]

Purpose of testing dataset

Testing dataset is an important component which serves as a final arbiter for the performance of the model. Testing set evaluates how well the model which is trained using training and validation datasets, generalizes to unseen data. In order to understand the model's practical effectiveness and its reliability for real world applications, this evaluation of the model using a testing set is important.

Testing dataset plays the following important roles-

- It plays an important role in final model evaluation. Testing dataset is used to assess the model's accuracy and performance after it has been trained and validated, using relevant metrics. This evaluation is necessary to determine whether the model can make accurate predictions on unseen data.
- Testing dataset plays an important role in verifying that the model generalizes well with the new data to ensure that it is capturing the underlying patterns in the new data instead of memorizing only the instances from the training set. This is important because it leads to overfitting if the model performs well on training and validation data but poorly on testing data.
- It plays an important role in providing a basis to compute the performance metrics like F1 score or other metrics. Computed metrics show the effectiveness and reliability of the model.

- In order to identify the robust model among different models based on GARCH, neural network or hybrid approaches, the testing dataset acts as a final benchmark to compare the performance of all the models.
- Testing dataset plays a crucial role in simulating the real world data as closely as possible and shows the performance of the model in practical applications.

Testing dataset is a vital component which provides the final evaluation for the model's performance. It helps in making informed decisions regarding model deployment. [2]

Comparison of Validation and Testing dataset

Typically, three main divisions of datasets are there, i.e., training set, validation set and testing set. Each dataset serves its own unique purpose for the model. In order to build robust models, it is crucial to understand the difference between validation and testing datasets.

Validation dataset	Testing dataset
It is primarily used for hyperparameter tuning.	It is primarily used for an unbiased evaluation of a model's performance after training and tuning.
It helps in comparing the performance of multiple models and selecting the best one.	It ensures the model's proper generalization to new and unseen data.
It prevents model's overfitting by evaluating its performance on validation sets.	It helps in computing final performance metrics.
It is used during the model development phase.	It is used when the model is fully trained and validated.
Evaluation metrics help in identifying the best hyperparameters.	Testing set provides final performance metrics which measure the model's effectiveness.

Table.1. Comparison of Validation vs Testi

Allocation of data to Training, Validation and Testing dataset

Macroeconomic and Financial data that we have since the beginning, is used for allocation. First of all we are combining the macro data with the financial data and after that we are dividing it into training (80%), validation (10%) and testing (10%) datasets.

First, the two datasets are being merged to form a dataframe and then two additional columns are added, one for the current crude oil price and next for the predicted or forecasted prices.

Validating the model and re-running Bayesian Network using hill climbing

In order to validate the model, hill climbing is used to run the Bayesian Network model on the discretized data. The data is discretized into binary values (1-increase and 0-decrease). At first the parameters of the hidden markov model are computed and then with an initial expert knowledge model, hill climbing is performed.

Finally, the Bayesian network model is fitted and the visualization of the resulting network is observed as follows-

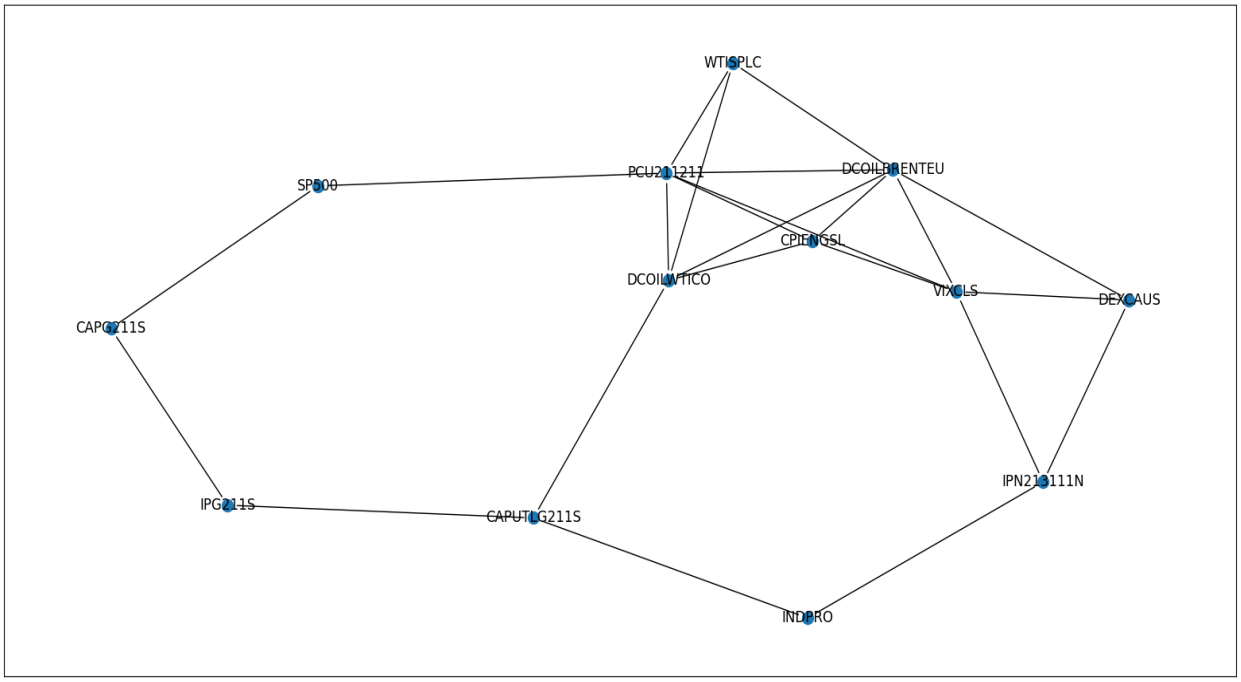

Fig.1. Bayesian Network model on the discretized data

The graph edges are observed as follows-

```
Graph edges: [('WTISPLC', 'PCU211211'), ('WTISPLC', 'DCOILBRENTEU'),
('WTISPLC', 'DCOILWTICO'), ('PCU211211', 'SP500'), ('PCU211211', 'CPIENGSL'),
('PCU211211', 'VIXCLS'), ('PCU211211', 'DCOILBRENTEU'), ('PCU211211',
'DCOILWTICO'), ('SP500', 'CAPG211S'), ('CPIENGSL', 'VIXCLS'), ('CPIENGSL',
'DCOILBRENTEU'), ('CPIENGSL', 'DCOILWTICO'), ('CAPG211S', 'IPG211S'),
('IPG211S', 'CAPUTLG211S'), ('CAPUTLG211S', 'INDPRO'), ('CAPUTLG211S',
'DCOILWTICO'), ('INDPRO', 'IPN213111N'), ('IPN213111N', 'DEXCAUS'),
('IPN213111N', 'VIXCLS'), ('DEXCAUS', 'VIXCLS'), ('DEXCAUS', 'DCOILBRENTEU'),
('VIXCLS', 'DCOILBRENTEU'), ('DCOILBRENTEU', 'DCOILWTICO')
```

Table.2. List of the edges of the Bayesian network model

In the plot above, it can be observed that the nodes in the plot represent the variables of the dataset whereas the edges represent the dependencies between them. It can be observed that the exchange rates, stock market and market volatility influences crude oil prices. Further, Brent crude oil futures (DCOILBRENTEU) and producer price index (PCU211211) can be categorized as the key variables since their nodes have many connections.

The fitted model is able to replicate the results of the paper as the models and parameters are correctly implemented according to the methodology provided in the paper. The learned structure of the Bayesian network is observed as somewhat different from the one given in the paper because the data used is different as we have used the macroeconomic and financial dataset. Further, it has been observed that Brent crude oil futures and producer price index can be the underlying factors driving crude oil prices. As our model has been fitted now so the inferences can be done using forecasts.

Interpretation of Results

In order to report the accuracy of the forecasted crude oil prices after successfully fitting the model to the data, first the testing will be done on validation dataset to know if any adjustments are required. After that, the testing will be done on the testing dataset to report the accuracy of the model.

First, the testing is done on the validation data and the data is discretized and the following results are observed for predicted or forecasted value and real values.

```
Predicted Value:
[2 0 2 2 2 2 0 2 0 2 2 0 2 0 2 2 0 2 2 2 0 2 2 2 2 2 2 2 0 0 2 0 0 2 2 2 2
 2 2 0 0 0 2 2 2 0 0 2 2 2 0 0 2 2 2 2 0 0 0 2 0 2 0 2 2 2 0 0 0 0 0 0 0 2
 0 2 2 2 0 0 2 2 0 0 0 0 2 2 2 2 0 2 2]
Real Value:
[0 1 0 1 1 1 1 0 1 0 1 1 0 1 0 1 1 0 1 1 1 0 1 1 1 1 1 1 1 0 0 1 0 0 1 1 1
 1 1 1 0 0 0 1 1 1 0 0 1 1 1 0 0 1 1 1 1 0 0 0 1 0 1 0 1 1 1 0 0 0 0 0 0 0
 1 0 1 1 1 0 0 1 1 0 0 0 0 1 1 1 1 0 1]
Error:
59.13978494623656
```

Table.3. Validation dataset Predicted vs Real values

Testing is done on the validation dataset and the error is computed as 59.139.

After that, the testing is done on the testing data and the data is discretized and the following results are observed for predicted or forecasted value and real values.

```
Predicted Value:
[2 1 1 1 2 1 2 2 1 1 1 1 1 1 1 2 1 1 1 2 1 2 1 1 2 2 1 1 1 1 2 2 1 2 1 2 1
 1 2 2 2 2 1 1 1 1 2 2 1 1 1 1 1 2 1 1 1 2 1 1 2 2 1 1 1 2 1 1 2 2 2 1 2 2
 1 2 2 1 2 2 1 1 1 2 2 2 1 1 1 1 2 2 1 2]
Real Value:
[2 0 1 1 1 0 1 0 0 1 1 1 1 1 1 1 0 1 1 1 0 1 0 1 1 0 0 1 1 1 1 0 0 1 0 1 0
 1 1 0 0 0 0 1 1 1 1 0 0 1 1 1 1 1 0 1 1 1 0 1 1 0 0 1 1 1 0 1 1 0 0 0 1 0
 0 1 0 0 1 0 0 1 1 1 0 0 0 1 1 1 1 0 0 1]
Error:
40.42553191489361
```

Table.4. Testing dataset Predicted vs Real values

Testing is done on the testing dataset and the error is computed as 40.425.

It is observed that the error of the testing dataset is less than the error of the validation dataset which makes the testing set more accurate. In order to display the results graphically, the following plot is observed for Bayesian model and crude oil prices-

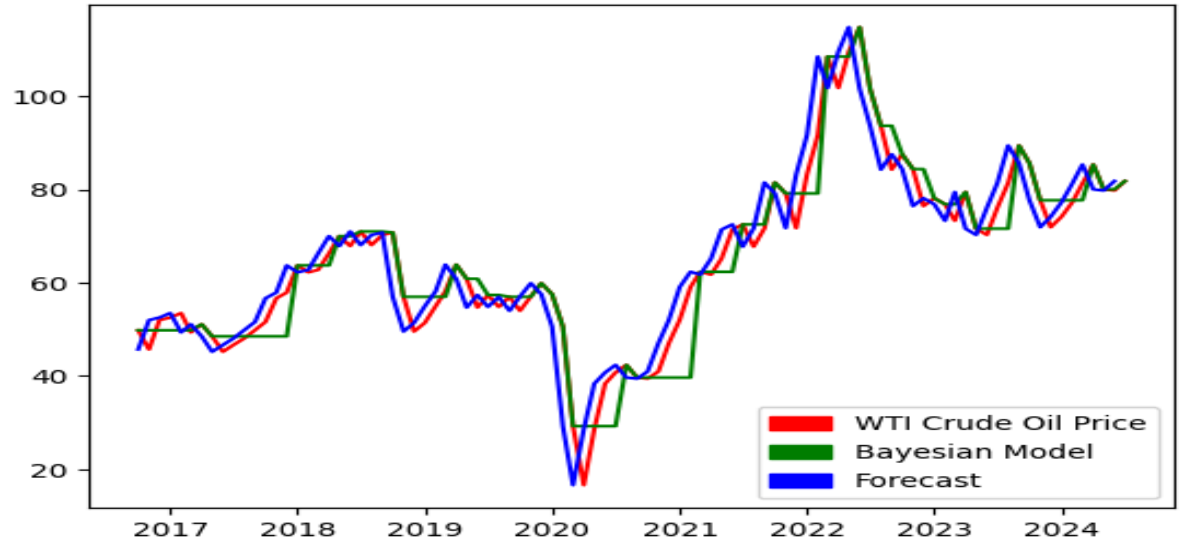

Fig.2. Plotting the WTI Crude Oil Price vs Bayesian Model's price vs Forecast price

From the above plot, a comparison can be made between the performance of the model and the forecast itself.

Contributions

The paper does:

1. It replaces the usual approach to forecasting time-series by GARCH group of models with the Bayesian Network Model and this is useful because this is how it brings conditional probabilities in the forecasting itself, while GARCH models do rely on autocorrelation and partial correlations between the data points without taking into consideration the possible sudden change of variable value or some additional conditional event that can bring new information into the table and thus evolve into a new unexpected pattern that was not involved previously and was not incorporated into the data beforehand. [3 pages: 51 … 65]
2. The paper is presenting a model that is not only suitable for trading strategies [3 pages: 34 and 43] but the idea itself and the implementation is suitable for vast majority of cases [3 pages: 35 and 43] where you have a priori data [3 pages: 36 … 39] and with time time some additional data comes in (a new event happens) that brings new information and does change the apriori data so the possible future outcome will change as well. This brings dynamics into forecasting that usual models cannot implement because mainly they are using predefined data.
3. Something is even better, that this Bayesian Network model [3 page: 13] and approach towards data is in practice a structural learning [3 page: 14] that needs input data with some conditional probabilities [3 page: 8] and brings out the most probable outcome based on the given data and its possibilities to change. It is very close to how the human brain works: it assesses the given data and based on the new incoming data makes decisions based on the new information so it is like an autonomous decision making process that is resembling the human like thinking process.
4. All this brings in the implementation of a dynamic event-driven systematic process [3 page: 67] of forecasting which takes into consideration a priori [3 page: 14] and aposteriori data [3 page: 18] and thus creates better and more accurate predictions because of the much more quantity and quality of data as well as its capability of dynamical change.
5. This is a novelty that combines Hidden Markov Model (a stochastic model) [3 page: 22] as inputs to Bayesian Network model [3 page: 8] (a probabilistic model) [3 page: 7]. The first one depends only on the current state while the second one depends on conditional dependencies of incoming and new data inputs and states. Such a combination of 2

totally different and opposite approaches, is novelty by itself and as if it brings the best of these 2 different worlds that makes the forecasting so much better than a normal and well known linear regression for example.

The author managed to accomplish his propositions because he did construct a Bayesian-based model that takes into consideration the dynamics of Oil market by implementing the macroeconomic data as well as the financial data dependencies and implication towards the Oil price. This means that has fulfilled his first goal of understanding the structure of macroeconomics and geopolitics of the Oil market and then building a model that incorporates all that.

His second accomplishment is the successful utilization of a lot of different data (different sources and different types of data). By downloading different time-series datasets from the federal reserve economic data site (FRED) and also the U.S. Energy Information Administration site (EIA), he achieves the goal of getting a reliable and high quality trustful open-data from governmental facilities (trustworthy sources). With this we can say that this data is relevant and very useful for building a price related model.

The further modeling, training, validation, testing, backtesting and stress testing procedures that were applied are in line with the model's performance validation. They were rigorously implemented and thus the outputs and conclusions that came out of it can be relied on with a strong conviction.

With all the achievements above an automated trading strategy was created that can be applied without specific a priori knowledge which makes it easy to apply and useful towards educational purposes, as well as some insights were provided for policy makers and trading practitioners. With all this the achievement of a practical application and usefulness was accomplished. So in other words said: this is a useful, educational and practical article that has great value as well as a lot of insights that could be applied into some different areas and cases.

All this is very important because it will open the gates for future and further analyses, predictions, education and many more applications. The author is leading the way for the furniture progress and development of many more new findings and insights in the vast majority of data and this is not closed in the realm of the financial markets, because this approach may be applied into different areas with different data, cases and problems that will bring new findings and statistical inference.

Discussion

This study can be categorized as a state of the art in forecasting because it incorporates the best techniques and methods from the both worlds: Hidden Markov Processes and Bayesian Networks. The statistical inference produced by the Bayesian Network is better than the ordinary existing prediction models because it incorporates a priori data and a posteriori data where some new information may come in and change a lot of things. It is also better about

cost efficiency because the conditional event that may or may not happen can make some data irrelevant and thus it can be not even taken into consideration because it does not bring new information or change anything so it can be skipped from the calculations. Imagine these conditional events as a switch that turns on and off the usage of some specific data. So in the case that an event does not happen means that this specific data is not useful and should not be used in the inference. This may drop a lot of unnecessary computations which will bring cost efficiency higher.

Furthermore it brings dynamics that are not applicable in the existing prediction models because they use the given data without the possibility of a new incoming data that may or may not change the model's behavior and thus modifying the final outcomes. Also a lot much more datasets can be involved in and some more insight could be derived. Especially with some future experimentations with the advanced algorithms like min-max hill climbing, that could bring a lot more datasets and thus make the predictions even more accurate (because of all the new information) or combine it with reinforcement learning that will bring an even better model's reaction to the dynamics.

REFERENCES:

1. Lee, S.B., Gui, X., Manquen, M., Hamilton, E.R. (2019). Use of Training, Validation, and Test Sets for Developing Automated Classifiers in Quantitative Ethnography. In: Eagan, B., Misfeldt, M., Siebert-Evenstone, A. (eds) Advances in Quantitative Ethnography. ICQE 2019. Communications in Computer and Information Science, vol 1112. Springer, Cham. https://doi.org/10.1007/978-3-030-33232-7_10
2. Khan, R. Deep Learning System and It's Automatic Testing: An Approach. Ann. Data. Sci. 10, 1019–1033 (2023). https://doi.org/10.1007/s40745-021-00361-w
3. Danish A. Alvi, Application of Probabilistic Graphical Models in Forecasting Crude Oil Price, Department of Computer Science at University College London, 2018