**GROUP WORK PROJECT #** 3 MScFE 660: RISK MANAGEMENT
**GROUP NUMBER:** 5638

| FULL LEGAL NAME | LOCATION (COUNTRY) | EMAIL ADDRESS | MARK X FOR ANY NON-CONTRIBUTING MEMBER |
|---|---|---|---|
| Shailza Virmani | India | virmanishailza@gmail.com | |
| Anubhav Mishra | India | anubhav0by0@gmail.com | |
| Zhe Zhang | Hong Kong | zhezhangcs@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).

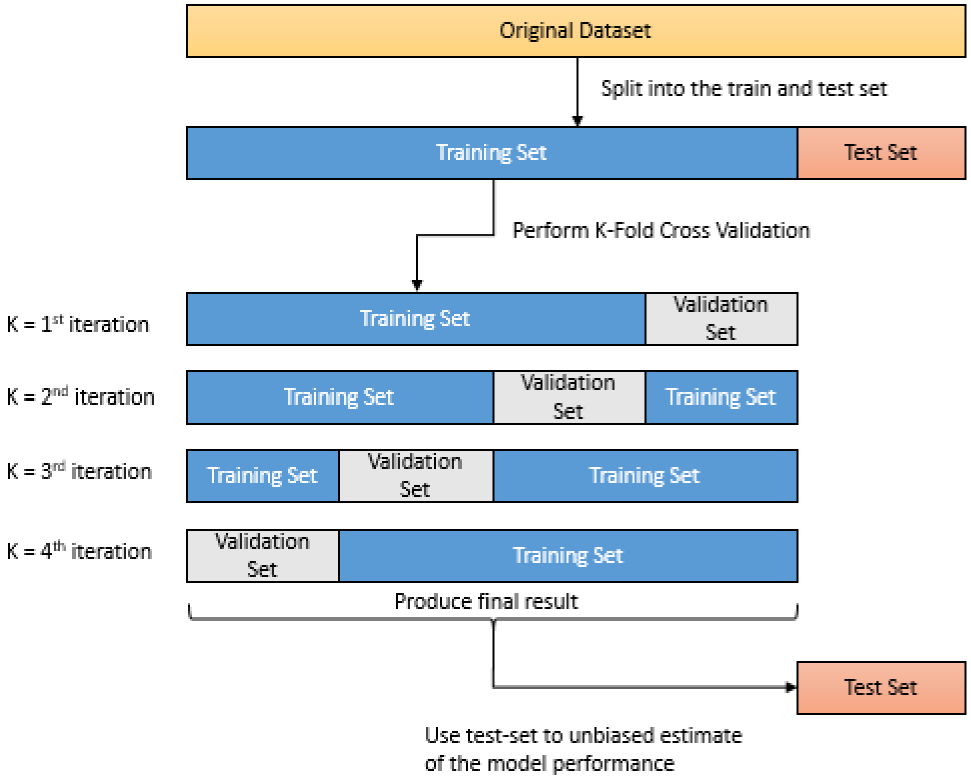
| | |
|---|---|
| **Team member 1** | **Shailza Virmani** |
| **Team member 2** | **Anubhav Mishra** |
| **Team member 3** | **Zhe Zhang** |

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

# Step 1-2

Post performing the task of preprocessing of the data which includes imputation of null values and removal of extreme scenarios which could alter the generalization capability of the model, we split the input data into the training and test set. This ensures that the data is trained on a specific set of data and tested for its accuracy or model metrics on the test set. The above methodology of splitting the data ensures that the model is not overfitting the data which makes it lose its generalization capabilities. The diagram below illustrates the methodology for any data manipulation before subjecting it to a generic machine learning model.



Generally speaking, the observations are split into the training, validation and testing data in the proportion of 80:10:10.

**Training Set** - The training set is the subset of observations on which the model is trained and the basis upon which the model is generated which can predict the outcome and aid the end user. The quality of the training data and accuracy metrics considered is an important driver in determining the quality of the model. Diversity of the data in terms of coverage, Low noise/variance, an evenly balanced dataset and large size of the training data with respect to the degree of freedom are factors which affect the performance of the model. A diverse dataset aids in improving the generalization capability of the model and prevents overfitting. Low noise in the training set reduces the incorrect learning of the model and thus making wrong outcomes. A large training set captures the complexity of the underlying data distribution and models learn more robust and generalizable patterns while a balanced training set ensures that the model is not biased towards the majority class. We allocate approximately 80% of the data to the training set.

**Validation Set -** The validation set helps us choose the best model and optimize it. Here the model is tested for overfitting if the model becomes extremely specific to the data on which it is trained. The model is tuned for parameters and hyper parameters so that there is balance between bias and variance. We allocate approximately 10% of the data to the validation set.

**Testing Set -** Once the model is trained and tuned, we utilize the data in this set to evaluate the performance of the model. The characteristics of the data in the test set is expected to be similar to the training set. We use various metrics to compare the score of the models and assess the performance when we roll out the model trained on the training set onto the test set. We allocate approximately 10% of the data to the test set.

As illustrated in Figure - 1, we apply cross validation on the dataset which splits the training data into the training and validation set where in each iteration of the cross validation we train the data on the training set and fine-tune the parameters on the validation set. The K-fold cross validation creates k different datasets from the training data and creates a model from each of the same. Thus, we perform the activity K times. The mean accuracy is then calculated as an average of each of these iterations. The final step would be to test the model obtained after the K iterations on the test set and that the results are captured. It should be kept in mind that the model should not be tuned after looking at the test results else it is a clear case of “data leakage”.

We have maintained a train, validation and test split in the ratio 80:10:10. We infer that the number of samples determines the dataset split ratio. Our intent with the model is to have a low bias and low variance in the model performance and given a large and diversified dataset helps us attain the same. It is thus of prime importance to understand what bias and variance imply in this scheme of things.

Bias: The simplifying assumptions made by the model to make the target function easier to learn are referred to as bias. Generally speaking, the linear algorithms given their simplicity learn the fastest, however, they suffer from Bias i.e. they have lower predictive performance on a complex problem given their simplifying assumption of linearity. A low bias suggests lesser assumptions about the form of the target function by the model and vice-versa. [4]

Variance: Variance is the amount that the estimate of change of the target function if different training data was used. The target function is estimated from the training data by a machine learning algorithm, so we should expect the algorithm to have some variance. Ideally, it should not change too much from one training dataset to the next, meaning that the algorithm is good at picking out the hidden underlying mapping between the inputs and the output variables. [4]

Machine learning algorithms that have a high variance are strongly over fitted with the training data. This means that the specificity of the training data has biased the model and types of parameters used to characterize the mapping function.

Low Variance: Suggests small changes to the estimate of the target function when the model is tested on a dataset different from the train set

High Variance: Suggests large changes to the estimate of the target function when the model is tested on a dataset different from the train set

We can take the following considerations into account when splitting the dataset:

The parameters of a machine learning model expand with the increasing the input features or dimensions of the input data. Thus, the complexity of the model scales up and we thus need a larger dataset which brings in diversity to the data. Every feature or attribute is like an additional dimension and adds to the complexity of the model.

a. The larger the number of parameters, which need to be adjusted, in a machine learning model, a larger validation dataset is needed and thus there is a possibility of a higher bias.
b. Smaller amounts of data or smaller datasets might result in larger variance in the machine learning models.
c. It is a good practice to validate the model after each epoch so as to ensure the model learns the most of the available data, especially if the cost of wrong output of the model is very high, e.g. incorrect diagnosis of cancer.

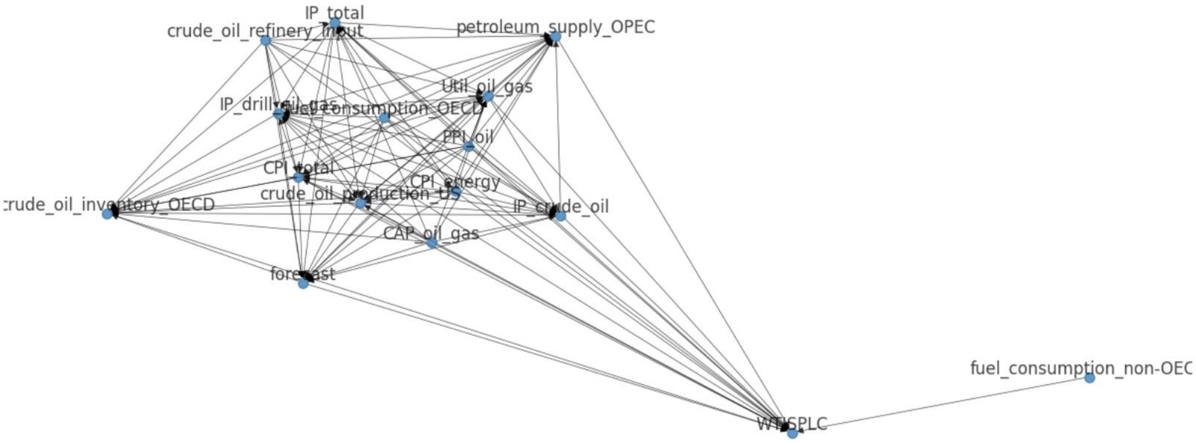
To bring in trustworthiness in the data, we follow the principles of:

1. Quantity - The volume of data should be large so that there is abundant data for model to learn the relationship between the input variables and the output target function
2. Quality - The data should be similar to actual or real-world circumstances
3. Diversity - To replicate the majority of the likely instances, machine learning algorithms should be trained on several folds of data
4. Overfitting - The separation of the validation and test set should be delineated else this will lead to overfitting which could be a perfect example of data leakage.

We use the F1 score for the performance metric of the model which is the metric derived from the Recall and Precision metric of the model. We combine different datasets of FRED and EIA into one data frame and then split them in the ratio 90:10 by slicing them. Then we perform a 10-fold cross validation on the data which further performs 10 iterations of the machine learning dividing the training data into the train set and validation set in the ratio 9:1.

## Step 3-5

We collect the data from 2003 to 2023 and re-run the Bayesian network using hill climbing and get the following network structure.



For the validation, we get the result:

Predicted Value:

```
[0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]

Real Value:

[0. 1. 2. 2. 1. 0. 0. 0. 0. 1. 2. 0. 0. 0. 0. 0. 1. 0. 0. 0. 1. 0. 1. 0.]

Error:  41.67%. Accuracy: 68.33%
```

For the test set, we get the result:

```
Predicted Value:

[0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]

Real Value:

[0. 0. 0. 1. 0. 1. 2. 2. 1. 0. 1. 2. 0. 1. 2. 1. 2. 1. 0. 0. 0. 1. 2. 0. 0.]

Test Error:  56.00%. Accuracy: 44%
```

We can find that the Bayesian model can only identify the bear state (state 0) correctly in the validation period and wrongly predict the stagnant state (state 1). In the test period, the model fails to identify the bull period (state 2) and stagnant state (state 1) as the validation period.

Through the above experiments, we find that we can replicate the results reported in the paper. However, due to some random seed issues, the state assignments are not stable, that is, a single data point can be classified into different states in different running trials. As a result, the accuracy of the predictions is also not stable.

## Step 6

By suggesting that views produced by the Black-Litterman model be swapped out for views obtained from the Bayesian model rather than views obtained from the EGARCH-M, the suggested work advances the discipline. This modification could increase the resilience and accuracy of the model by including Bayesian notions into the forecasting process.

The article offers a novel method for discretizing time-series data utilizing Hidden Markov Models and Belief Networks.

This novel strategy might offer new insights and improve the model's capacity for forecasting intricate correlations seen in the data.

The dissertation proposes an automated trading system that gains intelligence, learns from its trades, and makes better trading decisions, potentially giving commodities traders higher alpha. The emergence of algorithmic trading has the power to fundamentally alter the way that commodities markets decide.

It is claimed that the recommended trading strategy is almost entirely autonomous and only needs some expert knowledge to choose the appropriate datasets. With less reliance on human input throughout the decision-making process, this autonomous feature could improve trade operations.

The study uses Bayesian analysis for financial forecasting in an effort to enhance model performance in crude oil price prediction.

Using Bayesian principles could provide a more dependable framework for incorporating macroeconomic, microeconomic, and geopolitical factors into the forecasting process.

The research provides a potential technique that accounts for large geopolitical and macroeconomic changes: using an event-driven, systematic global macro strategy to generate superior returns. The ability to accurately predict and respond to uncertain geopolitical and economic shifts may be necessary for the successful implementation of this strategy, which can offer a thorough understanding of market dynamics and capitalize on significant events for profit generation. This requires sophisticated data analysis and risk management skills.

The goal of the article is to improve energy market models so that energy policymakers can make better decisions.

However, a deeper understanding of the fundamental makeup of the oil markets may help shape policy responses to economic events more successfully. However, the complexity of energy markets—which are influenced by a wide range of factors—makes it challenging to create accurate models that can withstand real-world market conditions.

A forward-thinking strategy is shown by the suggestion to improve quantitative commodities trading and raise alpha by incorporating research from several sectors into commodity markets. The study may offer investors and industry participants novel options by incorporating viewpoints and tactics from various disciplines. However, it takes a deep grasp of each subject, rigorous testing, risk assessment, and flexibility in response to shifting market conditions to successfully translate research findings from several domains into useful applications in commodities markets.

## Step 7

This work first discretizes the time series data, including the crude oil price data as well as the macro features related to oil production, using the Hidden Markov Model. It then feeds this discretized data into a Belief Network to study the causal relationships.

Utilizing the discretized data significantly speeds up the fitting procedure of the Belief Network. Additionally, it allows the Belief Network to better capture the underlying patterns and dynamics in the data without the negative impact of noise hidden in the raw data. The network learned by the model helps analysts understand the relationships between different factors. By revealing the directional influences and strengths of the connections between variables, the network model can guide the development of more sophisticated oil trading algorithms, such as utilizing the lead-lag effects between factors. It can also help inform the development of longer-term investment strategies.

## References

1. Alvi, Danish A. "Application of Probabilistic Graphical Models in Forecasting Crude Oil Price."*ArXiv*, 29 Apr. 2018, arxiv.org/abs/1804.10869.

2. Wikipedia Contributors. "Baum–Welch Algorithm." *Wikipedia*, Wikimedia Foundation, 17 Jan. 2022, en.wikipedia.org/wiki/Baum%E2%80%93Welch_algorithm.

3. "Hidden Markov Models - an Introduction." *QuantStart*, www.quantstart.com/articles/hidden-markov-models-an-introduction.

4. Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Chapter 7, "Model Assessment and Selection," pages 219-257, https://g.co/kgs/hkk5jrt.