Machine Learning project

Oil price prediction

Bogdana Kolic, Clemence Mottez,
Matheo Le Masson

Motivations

- Crucial for industries and decision-makers
    - Impact on the global economy, investment and trading strategies, supply and demand dynamics, risk management, energy policy formulation, and market forecasting
- Usual data: year and price
    - More comprehensive approach with 23 features
- Feature selection
    - determine what contribute to oil prices
    - deepen our understanding of the complex dynamics driving oil prices
    - provides us with enhanced predictive capabilities

Data

- How did we create our data set
  - Decided by ourselves what we thought could influence the price, asked domain experts
  - Sources: macrotrends.net, unctadstat.unctad.org, datasource.kapsarc.org, data.worldbank.org, tradingeconomics.com
  - Handle missing values
  - Normalization

$$x = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- Our dataset
  - 23 parameters, total of 24 columns
  - Monthly data from 1970 to 2022, total of 624 rows

Data visualization

- data analysis
- each variable distribution
- correlations

Models

- Which models did we study
- Hyperparameters
- Which one did we choose for feature selection (xgbRegressor + random forest ?)

Feature selection

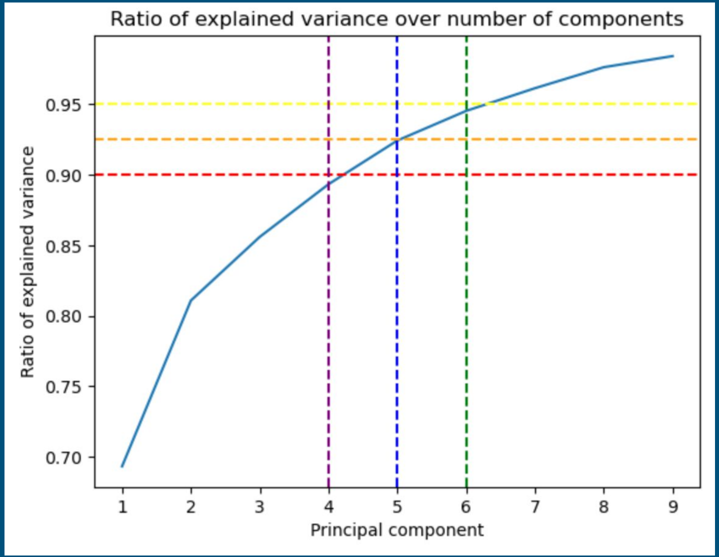
Why?

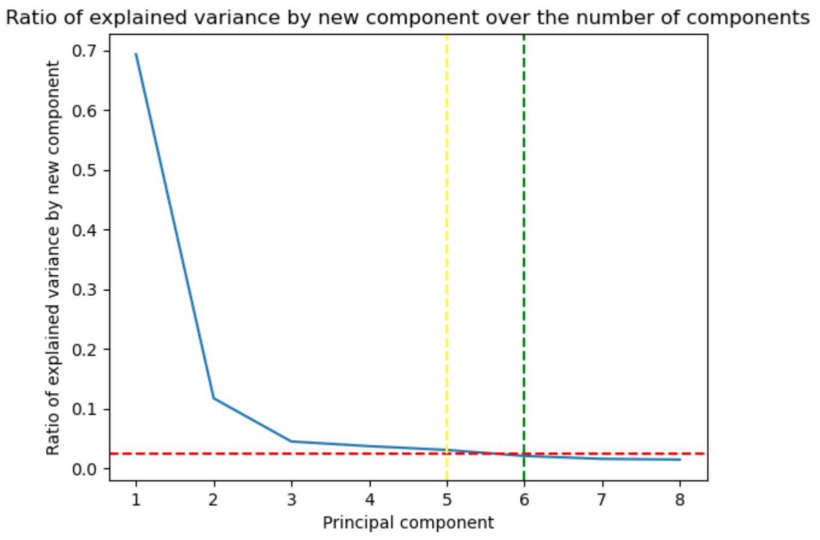
- Data 624x24!!!
    - Reduce computational cost

- Noise, redundant, irrelevant information have negative impact
    - Improve the performance of the model

- Capture essential dynamics of the oil market
    - Facilitate interpretation
    - What really predicts the price of oil?

PCA as an answer to: How many features should we keep?



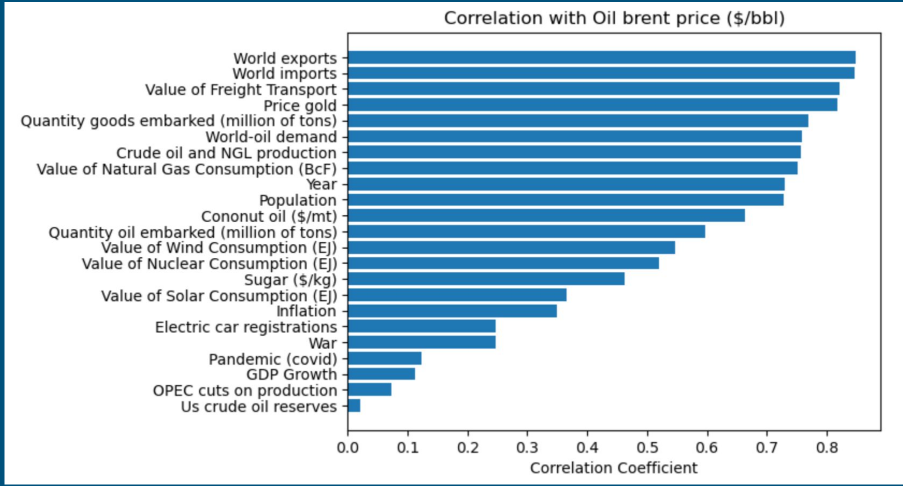


5 features explain 90% of variance

Components after the 5th explain less than 2.5% of variation

Filter models

- Computationally efficient
- Involve statistical measures such as correlation
  - Measures the linear relationship between each feature and the target variable



Embedded models

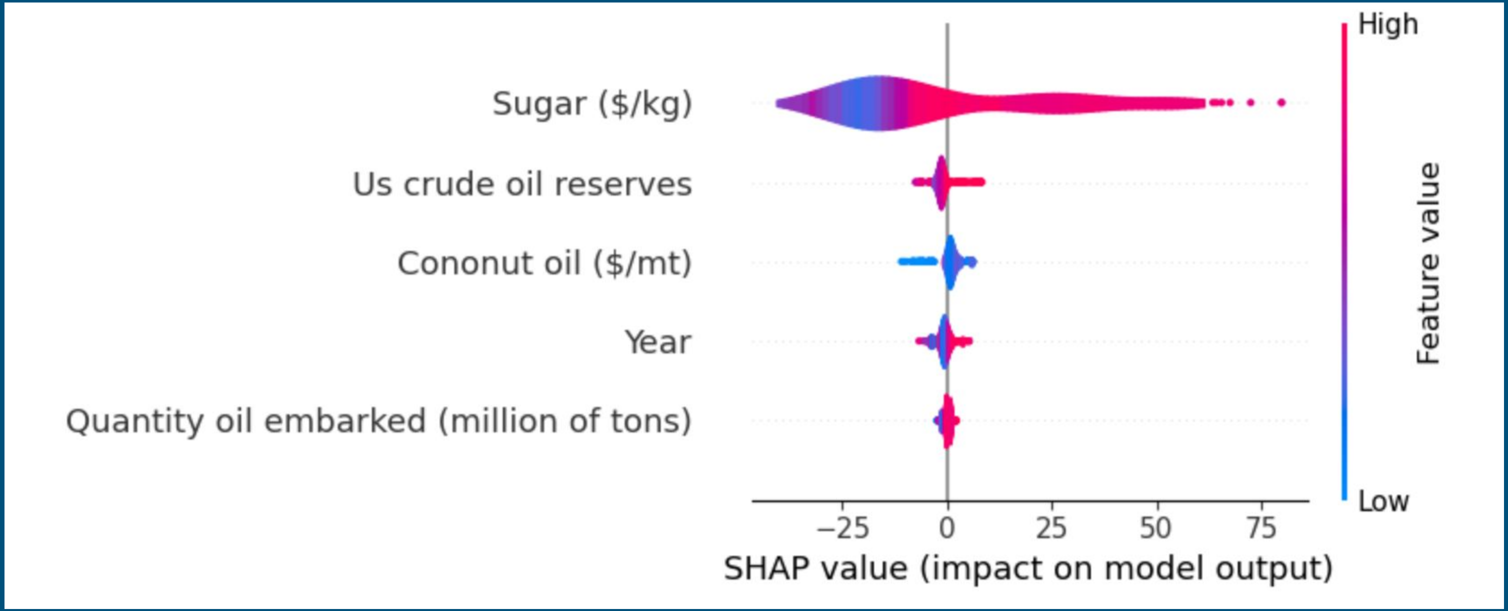
- Incorporates feature selection within the model training process
    - Optimizes both the model's performance and feature selection simultaneously
- Select relevant features based on their contribution to the model's accuracy

SHAP

- Interpretability tool
- Gain insights into feature importance and guide the feature selection process.
- Show global contribution
    - Computed for each feature and used to rank the importance of features
- Show local feature contribution
    - for each instance

Embedded models

- XGboost + SHAP



Wrapper models

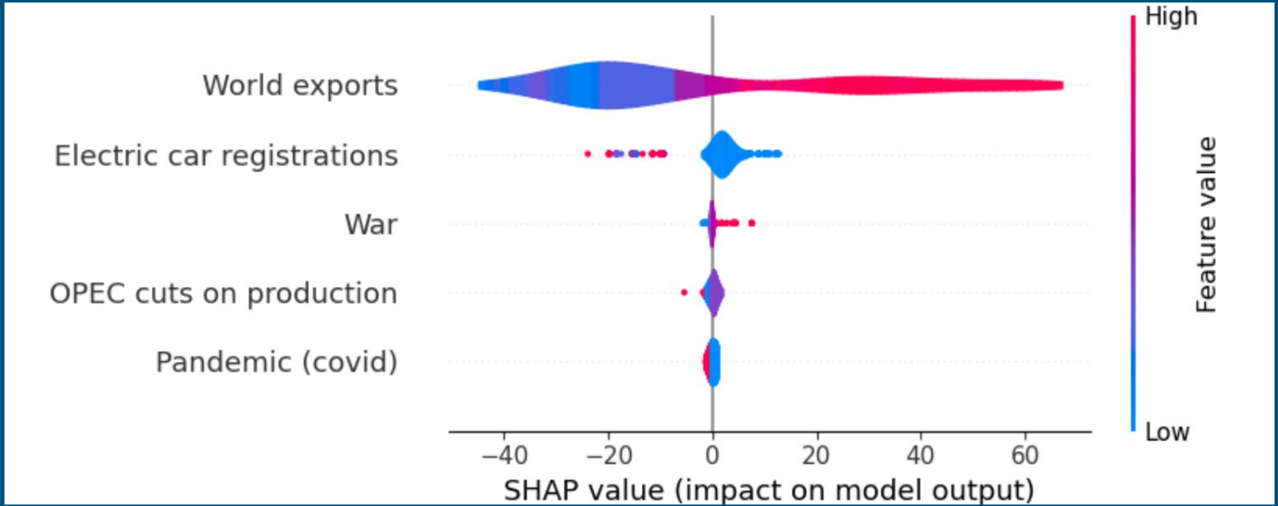
- Iteratively select and evaluate different subsets of features to find the subset that yields the best performance.
- Computationally expensive but accurate results

- Used Random Forest and XGB models

Wrapper models

Sequential Forward Selection

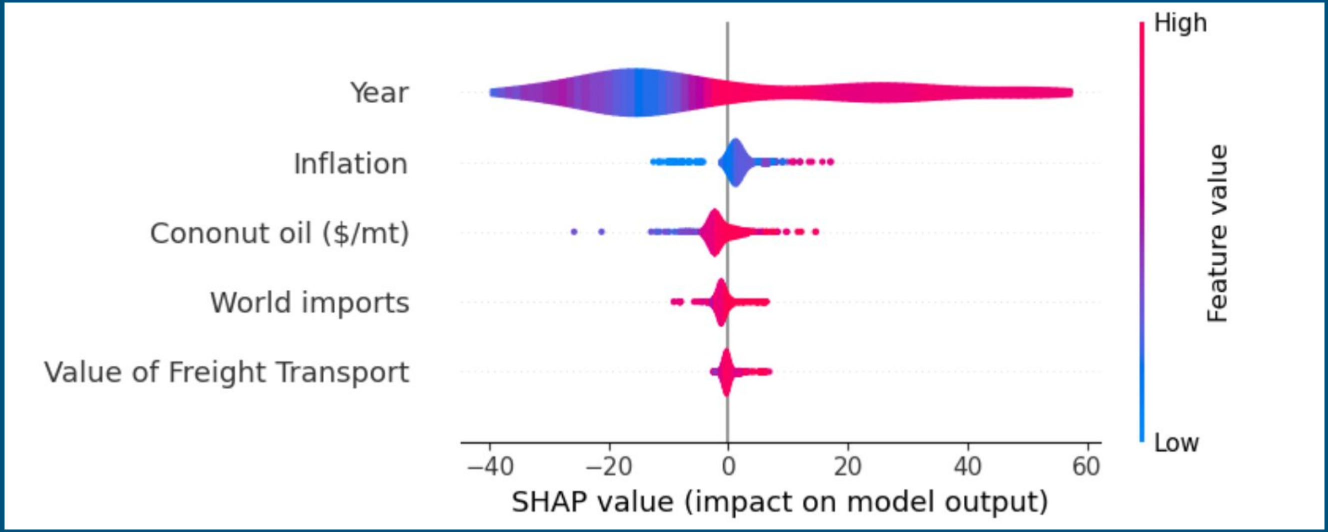
- Starts with an empty set of features
- Iteratively adds features based on the best-performing subset



Wrapper models

Recursive Feature Elimination (RFE)

- Starts with all features
- iteratively eliminates the least important feature



Wrapper models

Boruta

- A feature is important if it can do better than the best randomized feature

- Used eBoruta (extension of Boruta that already uses the SHAP importance)

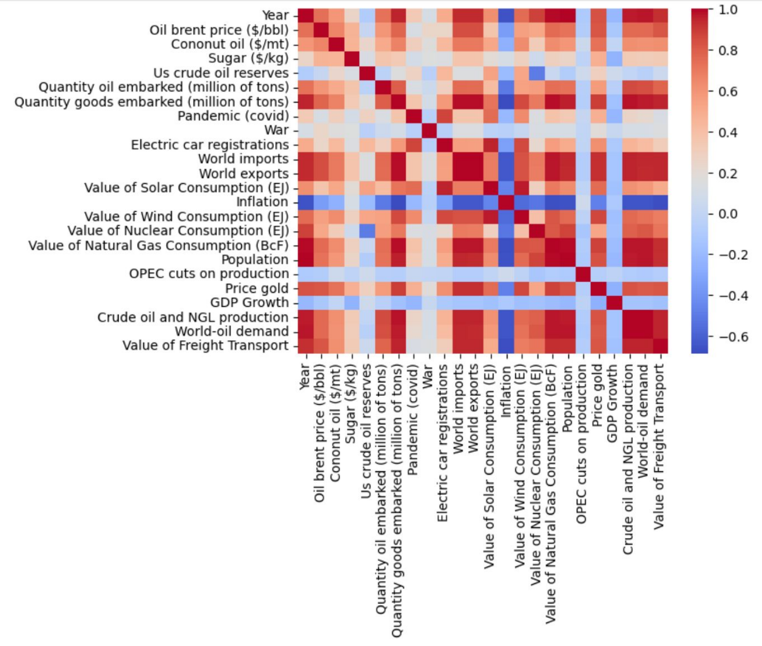
Feature	Importance
Year	20.353813
Inflation	2.827759
Cononut oil ($/mt)	2.487869
World imports	1.869873
Sugar ($/kg)	1.062382

Selection

- Year
- World imports
- World exports
- Inflation
- Price of Gold
- War
- OPEC cuts on production

Correlation

- Small selection so we don’t want too correlated variables
- Threshold at 0.95

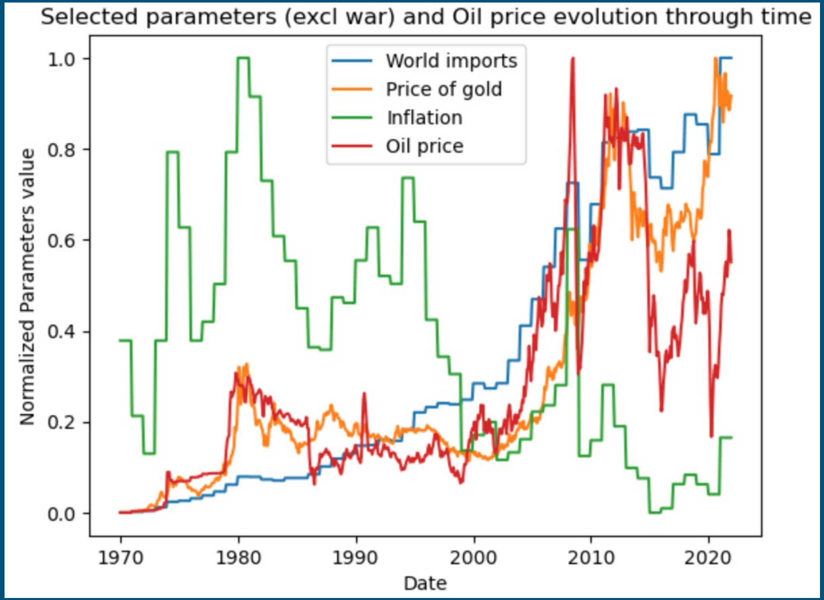


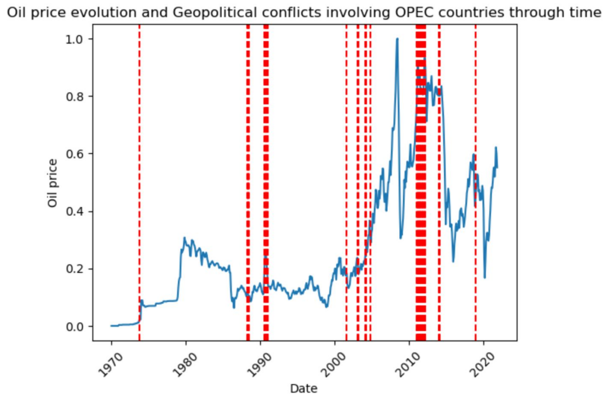
Result:

- World export and world imports

Final selection

- Year
- World imports
- ~~World exports~~ <- too correlated
- Inflation
- Price of Gold
- War
- ~~OPEC cuts on production~~ <- only 5 features selected





MSE

Data / Model	XGB Regressor	Decision Tree	Random Forest
Raw Data	**621**	643	680
Data with PCA	**253**	390	318
Selected Data	**372**	474	555

Conclusion

- 5 key features to predicting oil price
  - Year
  - World Imports
  - Inflation
  - Price of Gold
  - War
- Model with lowest MSE: XGB regressor
- Dimension-reduction reduces MSE

Limits - What could do after?

- Limits:
    - don’t know if causation or correlation
      for example price of gold certainly a correlation
- After:
    - Network model
    - Future analysis and predictions
      make predictions for future oil prices
      Incorporate new data as it becomes available
    - Optimize algorithm we used (for example SHAP is computationally very expensive)
    - New model: RNNs with LSTM to avoid Vanishing Gradient problem

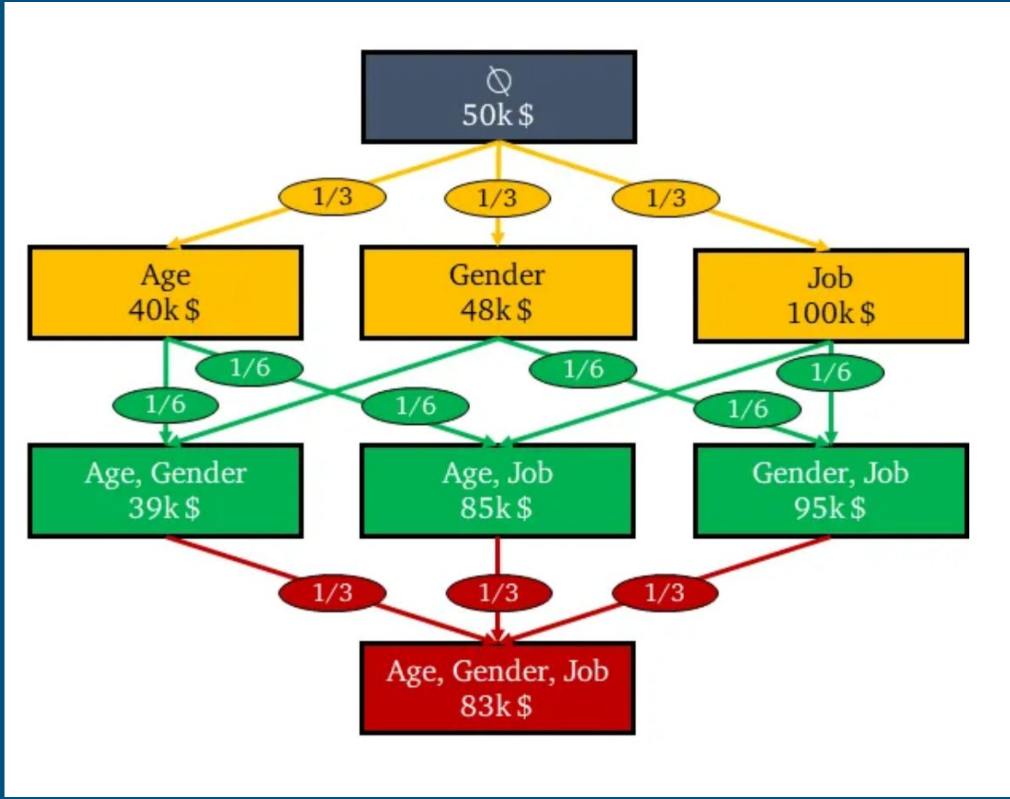
Network model

- Much better accuracy
    - MSE = 220 (compared to XGB = 620)
- Relatively simple:
    - 3 Hidden Dense layers with Relu activation
    - Dropout regularization to avoid overfitting

Other info for questions

- slides to explain the different models we used
- slides with term explanations, …

SHAP values



∅
50k $
1/3
1/3
1/3
Age
40k $
Gender
48k $
Job
100k $
1/6
1/6
1/6
1/6
1/6
1/6
Age, Gender
39k $
Age, Job
85k $
Gender, Job
95k $
1/3
1/3
1/3
Age, Gender, Job
83k $