Using machine learning to forecast price reaction of crude oil over a short time frame.

Shane Flynn
National College of Ireland
Dublin, Ireland
x18122957@student.ncirl.ie

Martin Mohan
National College of Ireland
Dublin, Ireland
x18191339@student.ncirl.ie

Declan Moore
National College of Ireland
Dublin, Ireland
x18150713@student.ncirl.ie

Bolu Obitayo
National College of Ireland
Dublin, Ireland
x16138821@student.ncirl.ie

***Abstract*—Like all products, price of oil is mainly decided by the basic economics of supply and demand. As majority of the world/industries is heavily reliant on oil the group have decided to predict the price of oil exploring various variables that influences the movement in price. The goal of the project is to use techniques learnt in the Advanced Data Mining module to accurately predict the prices of Oil at a particular time based on historic data. i.e. build a model that will predict the short-term price change in crude oil that occurs after the release of the new supply data. In summary the motivation of the paper is as described in the title using machine learning to forecast price reaction of crude oil over a short time frame.**

***Index Terms*—oil price, Machine Learning, time series analyis**

I. Related Work/Lit Review

Academic work has shown that total inventories of crude oil and refined products largely explain crude oil prices. Total inventory of petroleum inventory levels are a measure of balance between supply and demand [1]. In the short run supply and demand are inelastic. Inventories therefore reflect market fundamentals and their pressure on price in the short run [2].

It has been demonstrated that the crude oil market has short-term inefficient behaviour which becomes efficient in the long-term [3]. It has been discovered that unexpected inventory changes rather than the actual inventory changes that affect crude oil prices. This is the basis for our choice of using the difference between the survey of expected inventory changes and actual changes, rather than the absolute inventory change itself. Bu discovered that survey data has greater predictive power of the EIA inventory report data than the API report data [2].

Price movements based market fundamentals is one of the least understood factors of the price discovery process [2]. Papers attempting to forecast the price of crude oil tend to work on a time period of days and months. Baruník and Malinská used a dynamic Nelson–Siegel model to forecast oil prices using a generalized regression framework based on neural networks with success, however it was over 1-month-, 3-month-, 6-month- and 12-month-ahead forecasts [4]. Pan et all used ANN for modeling prediction over time horizons of 1-3 days [5]. Other papers attempt to simply discern the direction of price movement, rather than determine a quantifiable price movement [6]. The literature for intraday prediction is limited.

Additionally, despite the demonstrated importance of the inventory changes on price, the literature regarding the impact of weekly crude oil inventory reports on crude oil prices and their volatility is also limited [2]. Our research is novel in that it examines both of these areas and attempts to build a model for the effect of inventories and short-term price inefficiencies in intraday oil prices.

A recent review of machine learning in energy economics and finance provides a list of ML techniques. It states ”Papers dealing with forecasting crude oil prices are predominantly based on advanced and hybrid versions of ANNs and to a lesser degree of SVM models. Also, combining multiple methods (ensemble approach) has become more common in recent years.” Due to the many techniques available an ensemble technique was tried based on [7]. Using the library caretEnsemble enabled the comparison of a large number of models to find which was the best predictor of Price Change. We found that glmboost to best method followed closely by linear regression.

II. Data Mining Methodology

Our approach to the problem is based on CRISP-DM ...

1) Business Understanding II-A
2) Data Understanding II-B
3) Data Preparation II-C
4) Modeling II-E
5) Evaluation II-D
6) Deployment II-F

A. Business Understanding

The supply of crude oil is announced every Wednesday at 15:30. This has an immediate effect on price which is revised by 15:31. Can we predict the price change from 15:30 to 15:31 based on knowledge of crude supply change and several other supply factors

B. Data Understanding

The U.S. Energy Information Administration (EIA) provides a Weekly Petroleum Status Report (WPSR) [8]. This reports timely information on supply and selected prices of crude oil and principal petroleum products. The data are released at 10:30 a.m. Eastern Standard Time (EST) each Wednesday. The domestic and international oil market adjust rapidly to this information release. The goal is to build a model that will predict the short-term price change crude oil that occurs after the release of the new supply data. The model will be using the West Texas Intermediate (WTI) crude oil benchmark. Due to the large number of global inputs into the oil market it is difficult to model long-term price changes. By focusing on the immediate short-term market impact of the weekly supply data we hope to model an accurate prediction of the price changes to supply changes.

The weekly petroleum inventory report from the U.S. Energy Information Administration (EIA) has a major impact on crude oil prices in the US (Sundria and Smith) [9]. A survey of analysts predictions for the inventories sets market expectations. Any large deviation from this expected number can cause a major price change as the market adjusts to the new level of supply. Higher than expected inventories generally lead to lower prices and vice versa. For example, for the report on May 1st of this year; analysts' expectations were for an increase of increase of 1.5 million barrels in inventory levels. Instead crude inventories rose 9.9 million barrels to 470.6 million according to the report. Oil prices extended losses after the bearish report, with West Texas Intermediate crude dropping by more than $1 to a session low of $62.89 a barrel (Sundria and Smith) [8]. A Weekly Statistical Bulletin is released by American Petroleum Institute is released on Tuesdays (API). It is a similar survey to that released by the EIA but it is voluntary, unlike the government report, and covers about 90% of the available data. The API data is released the evening before the EIA report and it's seen as a prelude and predictor of the EIA data. However the EIA is seen by traders and analysts as more accurate than the API's (Palmer) [9].

In addition to reporting on crude inventories, both the API and EIA reports also report the levels of refined crude products. The inventory of levels of these refined products also has an impact on crude oil prices as it impacts on future demand for crude products. If there is a high inventory of distillates, then less crude is needed in the short term future. When the EIA report is released on Wednesdays, it is the difference between market expectations, set by the analysts poll, and the API, and the actual report that's delivered by the EPA that drives crude oil prices immediately following the report. We have obtained weekly report data from the API and EIA websites dating back to 2016, and publically available market price data for West Texas Intermediate crude on a short term time frame following each report.

C. Data Preparation

Using the Data Sources listed in section II-B we were able to obtain 6 files in xlsx format and csv format. We wished to predict price reaction over a short time based on our knowledge of supply change.

The dependent variable is Price and the independent variables are supply of Crude, Distillates, Gas , Cushing and Tuesday's API figure.

The data from each of the 5 files was read into data frames using R see table.I. The dependent variable Price difference (PriceDiff)was calculated by subtracting the Price at 15:31 from the price at 15:30 using information from the file 'DOE Prices.xlsx'. The independent variables Crude, Distillates,Gas, Cushing and API were calculated by subtracting the values actual supply-survey median from the files.

All files had a value utc_datetime_scheduled. As a first step files were merged using a modification of this time value (the time part was removed and only date retained and for API which is taken on Tuesday one day was added).

After all the merge the remaining values were checked using original time stamp to insure all independent variables were recorded before 15:30:01 each Wednesday.

The data was checked for normality see fig 2 .The diagrams and qq plots show reasonable normality. Checks for multicollinearity were performed see fig 1 and are also reasonable. Outlier were searched for and found by inspecting the plot for linear regression residuals and leverages shown in figure 4. The value of 131 for PriceDiff on 2018-06-06 looked like an outlier.This class was initially removed but then replaced as large outliers are the best way to acheive profit see section II-F

Field ref [10] page 274 recommends a minimum number

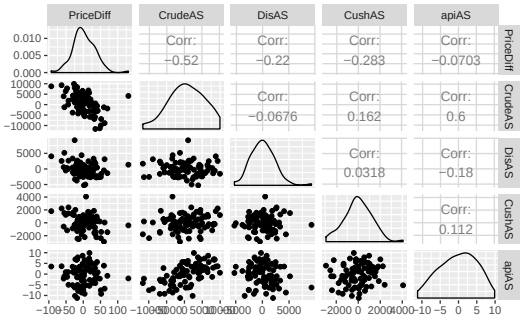


Fig. 1. Multicollinearity Plist

of samples for linear regression is 50+8k, where k is the number of predictors (so for 4 predictors 82 samples are required). 86 sample were available to us after cleaning data nrow(OilPrice)=86.

The main cleaning required was to remove values which had prediction dates after 15:31 on Wednesday (to avoid predicting after the fact)

As GasAS was not significantly significant it was not used for predictions. A list of dependent and independent variables is

Crude	Supply (actual - survey median)
Distillates	Supply (actual - survey median)
Gas	Supply (actual - survey median)
Cushing	Supply (actual - survey median)
API (Tue)	Supply (actual - survey median)
PriceDiff	**Crude Price at 15:31 - Crude Price at 15:30**

TABLE I
DEPENDENT (PRICE) AND INDEPENDENT VARIABLES

datetime	CrudeAS	DisAS	GasAS	apiAS	CushAS	PriceDi
2018-07-25	-3147	-651.0	-267	-1236.0	-3.16	-4
2018-08-01	6803	2483.0	-211	-536.0	5.59	15
2018-08-8	1649	430.0	748	4800.0	-6	-7

TABLE II
EXTRACT OF OILDATA TABLE

shown in table I an extract from the final table is shown in table II

D. Evaluation

Regression testing makes some strong assumptions about the data. These assumptions are not as important for numeric forecasting, as the model's worth is not based upon whether it truly captures the underlying process-we simply care about the accuracy of its predictions ref [11] page 197.
The detection of outliers was the most important thing. The initial data cleanup also used plots of residuals 1 to find outliers and acted on all data but subsequent modeling would split data 75% training and 25% test data.

E. Modeling

1) Linear regression: As discussed in the previous section II-C initial linear regression performed on the data and is listed below II-E1 .The results of the regression were also used to look for residuals 4. As a result of preparation some classes were removed because the dates were after prediction date and GasAS was removed as it was not statistically significant.

```
Call:
lm(formula = PriceDiff ~ CrudeAS + DisAS + apiAS + CushAS + GasAS,
    data = OilPrice)

Residuals:
    Min      1Q  Median      3Q     Max
-45.383 -12.980  -2.838  12.687 153.386

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.3932241  2.8112544   1.918  0.05862 .
CrudeAS     -0.0047169  0.0007223  -6.531 5.57e-09 ***
DisAS       -0.0027500  0.0013034  -2.110  0.03799 *
apiAS        2.2006837  0.7261164   3.031  0.00329 **
CushAS      -0.0053769  0.0022608  -2.378  0.01978 *
GasAS       -0.0004945  0.0011591  -0.427  0.67083
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25.66 on 80 degrees of freedom
Multiple R-squared:  0.4447,Adjusted R-squared:   0.41
F-statistic: 12.81 on 5 and 80 DF,  p-value: 3.741e-09
```

As b-values are easier to look at using QuantPsyc value [10] page [283] we decided to look check these using the lm.beta(). CrudeAs made the most significant contribution followed by apiAS,DisAS. >QuantPsyc::lm.beta(results)

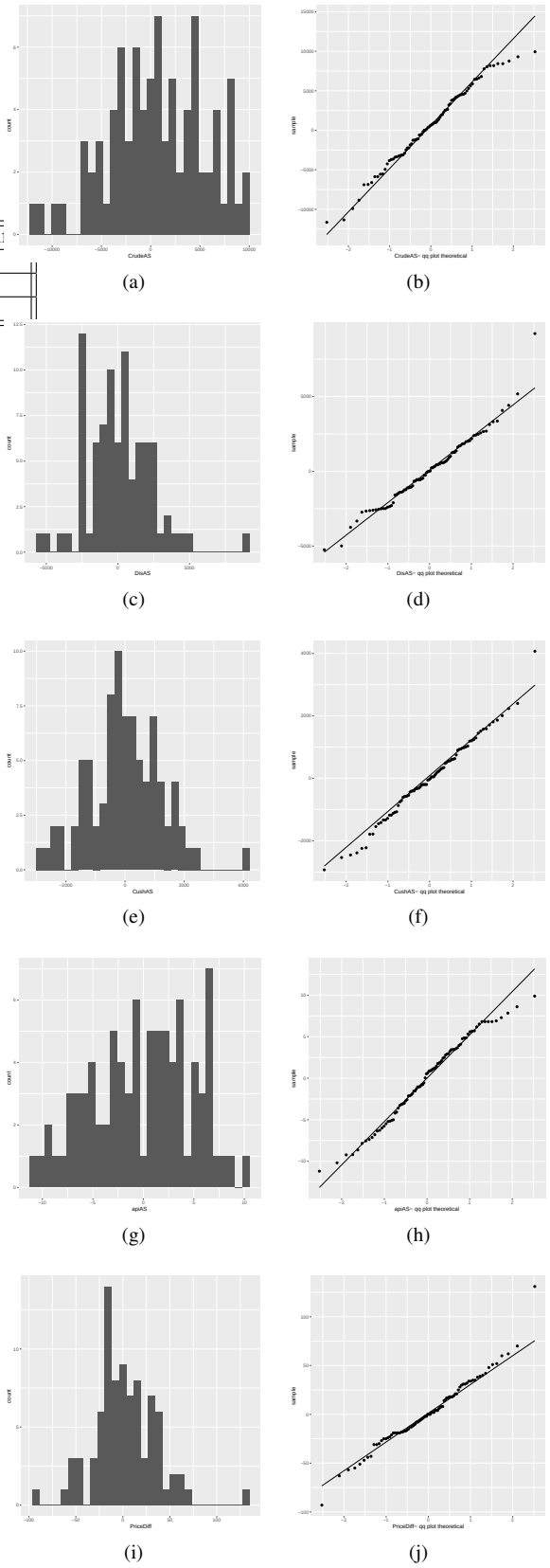


Fig. 2. Histograms / qqplots for (a,b) CrudeAS (c,d) DisAS (e,f) CushAS (g,h) apiAS(i,j) PriceDiff and apiAS(k,l)

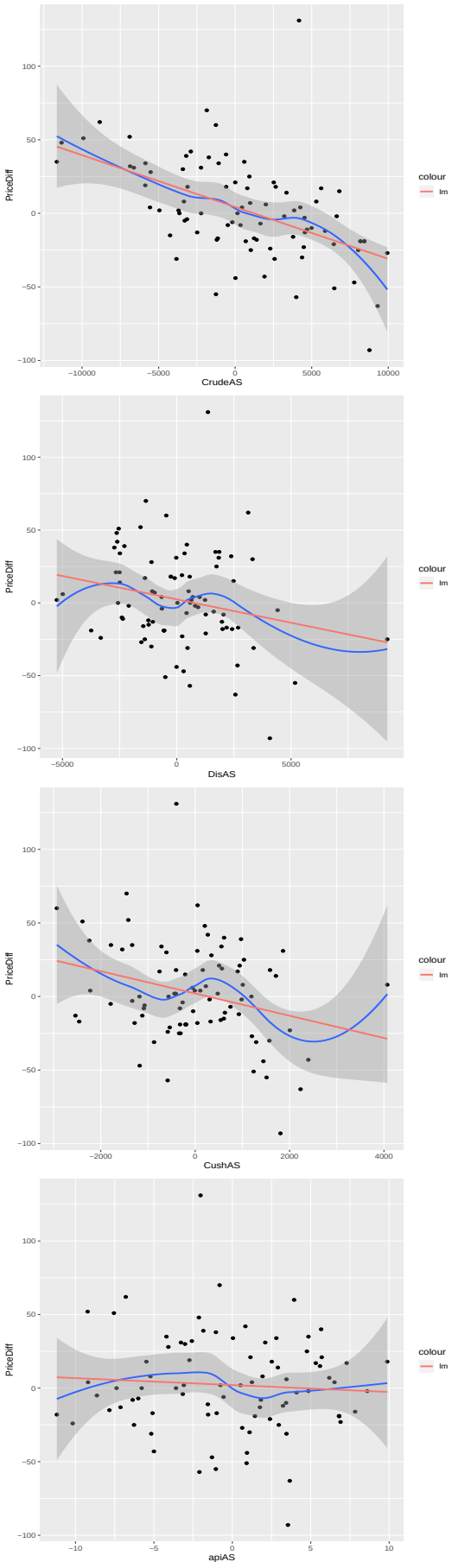
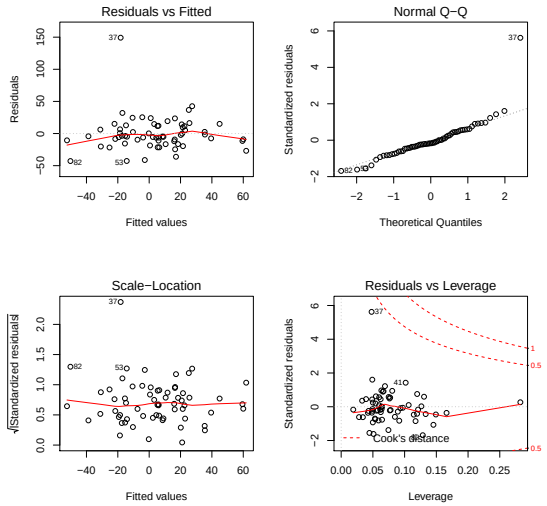


Fig. 3. Scatter Plots: PriceDiff vs Predictors



(a)

Fig. 4. Residual Plots to look for outliers

1) CrudeAS -0.8794529
2) apiAS 0.5577526
3) DisAS -0.2524455
4) CushAS_1 -0.2113192

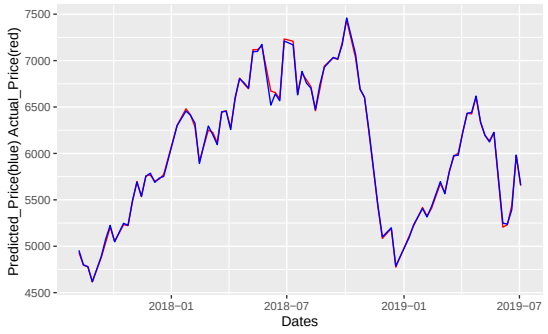
A graph of predicted price vs actual prices shows good accuracy fig 5 . The AUC score for the ROC was also calulated as 0.999 which is classified as outstanding but was not a good value for model comparison. The best measures for model comparison were the MAE **Mean Absolute error** and RMSE **Root Mean Square Error**. A representaion of MAE for linear regression is in figure 5 (b). Using these numeric values we could compare the performance of several models later II-E

In summary:Multiple regression was used to assess the ability of 4 control variables (CrudeAS,DisAS, CushAS, apiAS) to predict the price change in Crude oil after the announcemount of Crude oil supply at 15:30 each Wednesday. Preliminary analyses were conducted too ensure no violations of multicollinearity, homoscedasticity, normality and linearity. The independent variables showed the predictors CrudeAS, DisAS, CushAS and apiAS made a significant contribution to the model. GasAS did not make a significant contribution (p>0.05). R^2 of 0.44 means predictors account for 44% of the variance and as adjusted R^2 is very similar 0.41 it means the model should generalize well.

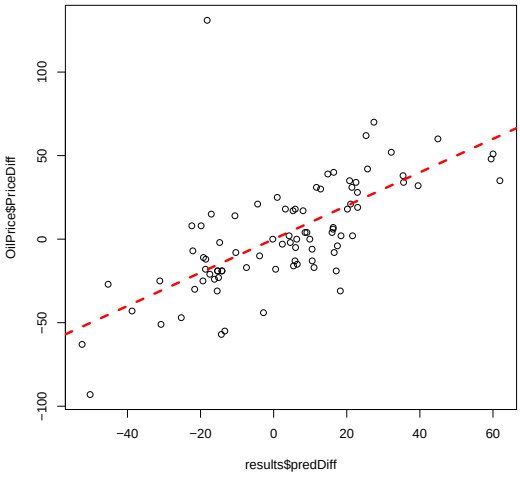
2) Other Models: Several models were then created using the caretEnsemble package. In this case there was a training data (75% of the OilData). This followed the methodology in [12]. Models were compared with training data using MAE and RMSE as plotted in fig 6. The results were very similar.

Four of the most promising models were selected for further invesigation fig 6.

The summary of the values shown below in fig 6 and the values for RMSE are shown in II-E2



(a)



(b)

Fig. 5. (a) Actual Price Difference (b) How residuals differ Predicted v Actual

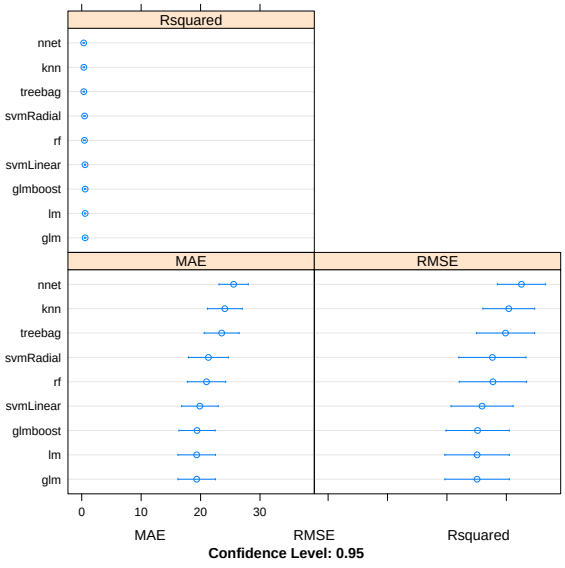


Fig. 6. Comparison of machine learning methods

```
Call:
summary.resamples(object = results)

Models: rf, lm, glm, glmboost, nnet, treebag, svmLinear, knn, svmRadial
Number of resamples: 30

MAE
             Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
rf       8.359378 15.14670 18.89606 20.99299 24.77963 43.19010    0
lm       8.235153 14.32619 16.43496 19.35096 20.16490 46.95671    0
glm      8.235153 14.32619 16.43496 19.35096 20.16490 46.95671    0
glmboost 8.687806 14.83273 16.57444 19.40600 20.52960 45.24156    0
nnet    16.000000 19.87500 25.30001 25.56469 29.42500 42.33334    0
treebag 11.828246 18.56296 21.34727 23.54974 27.22052 44.63351    0
svmLinear 11.397727 15.25742 16.81178 19.86826 19.99069 49.76605    0
knn     12.666667 18.60648 22.03671 24.08183 28.09722 45.59259    0
svmRadial 13.198699 15.99624 18.37335 21.31734 22.04549 50.19357    0

RMSE
             Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
rf       12.04915 18.39461 22.67460 27.74245 33.20636 69.89454    0
lm       10.83380 17.22761 19.74941 25.07741 26.11927 70.03333    0
glm      10.83380 17.22761 19.74941 25.07741 26.11927 70.03333    0
glmboost 11.58358 17.58181 20.70210 25.16064 26.63876 68.90106    0
nnet     19.16097 25.12344 30.97427 32.54024 36.18905 60.72341    0
treebag  13.57756 21.78810 26.30034 29.85689 31.38391 66.92074    0
svmLinear 15.70667 18.73733 21.85238 25.90922 24.97542 71.52552    0
knn      15.13254 23.06182 28.39399 30.40722 32.79384 62.43364    0
svmRadial 15.34042 19.30431 22.66546 27.65336 27.62141 74.27798    0

Rsquared
                Min.    1st Qu.    Median      Mean   3rd Qu.      Max. NA's
rf       8.196761e-04 0.21648981 0.4552131 0.4590341 0.7337479 0.9273232    0
lm       8.377001e-03 0.35437076 0.6608985 0.5758328 0.8245231 0.9800979    0
glm      8.377001e-03 0.35437076 0.6608985 0.5758328 0.8245231 0.9800979    0
glmboost 1.101175e-02 0.34963693 0.6677009 0.5732709 0.8158741 0.9744643    0
nnet     3.263974e-03 0.08471977 0.2951106 0.3230892 0.5368450 0.7744715    3
treebag  2.399133e-03 0.08660289 0.3370458 0.3460818 0.5774687 0.9099319    0
svmLinear 1.354415e-02 0.39865491 0.6256106 0.5573156 0.7870676 0.9473076    0
knn      2.780363e-06 0.09876305 0.2946627 0.3556556 0.5220718 0.9818089    0
svmRadial 2.573978e-02 0.22788272 0.5899971 0.4832384 0.6749112 0.9538778    0
```

The results of the models were fed into the caret functions caretEnsemble(models) ensemble_1 and caretStack(models) ensemble_2. These methods try to improve the result using ensemble methods.

The two ensemble methods and the four original methods were run against the test data using the predict.train function. The predicted results were compared against the actual results using the RMSE function using the test data. The results are listed below in II-E2.

1) XGBL 19.89
2) ensemble_2 19.94
3) LM 20.43
4) RF 21.37
5) SVM 21.48
6) ensemble_1 23.79

The result show that all the models have similar results and the ensemble methods to not improve the accuracy.

F. Deployment

The models produced can be used to predict the price of crude oil at 15:31 on Wednesday immediately after the Supply report received at 15:30. How can the model be used to maximize profit.

The graph in fig 5 (a) shows how closely the predicted price of oil follows the actual price. The simplest strategy is buy oil if a positive price difference is predicted and sell oil if a negative price difference is foreseen, however transaction costs (brokers fees etc.) and prediction error must be considered.

Profit = Price Difference - costs

Examining figure 5 (a) and the data shows that the price difference varies from a minimum of 0 (several times) to a maximum 131 on 2018-06-06, when the price rose from 6541 to 6672 dollars (2%)

If the price difference is 0 or very low then any profit would be consumed by transaction costs. In order to make a profit it should be estimated which price difference predicted would be enough to trigger a buy or sell?
Initially the 131 price difference was removed as it was thought to be and outlier but the large price differences are the ones which will generate a profit.
The most profit can be made from outliers but outliers are also the hardest to predict.

III. Evaluation/results

Initially Multiple regression was used to assess the ability of 4 control variables (CrudeAS,DisAS, GasAS, apiAS) to predict the price change in Crude oil after the announcemount of Crude oil supply at 15:30 each Wednesday. Preliminary analyses were conducted too ensure no violations of multicollinerity, homoscedasticity, normality and linearity. The independent variables showed the predictors CrudeAS, apiAS, CushAS and DisAS made a significant contribution to the model. GasAS did not make a significant contribution ($p>0.05$). The information gained above was used to decide on the data to be used for modelling. Four models for machine learning were compared to see which was the best predictor of Price change 6. The top 4 models were then selected. Two ensemble methods were also investigated but these models proved to be ineffective.
The difference between the top results were very close. The best model with was glmboost(xgbl) but only by a small margin.
Although all the models generally perform well a future investigation may be to find a model which is tuned to predict the important outliers. Other predictors may be also improve the model such as measuring sentiment on twitter using using phrases "tensions in the middle east".
In order to generate a profit it would be necessary to calculate the percentage price difference necessary to trigger a buy of sell of oil. This would take into account slippage costs and the error margin for prediction calculated in this paper.
It would be also useful to check the results are still valid with an Oil Price taken at a slightly earlier as 1 minute is little time to make a deciscion (unless the process is totally automated.)

IV. Conclusion and Future Work

Several models were made to predict oil price over a short time period. This time period was between 15:30 and 15:31 on Wednesday whenever when a supply report was released which affects oil prices immediately. Glmboost proved slightly better than the other models for price prediction with an RMSE of 19.89 dollars.
As price jumps were shown to range from 0 to 131 dollars this margin of error may be sufficiently small to generate a profit when big price jumps are predicted. Unfortunately large price differences are mainly the outliers and do not predict as well as the majority of price changes. The largest price jump would have generated a 2% profit without taking into account transaction costs.
Future work would be to calculate a percentage price difference which will trigger a buy or sell of oil. This would take into account slippage and margin of error for prediction. In addition it may be possible to create a better model which is more tuned to outlier performance where the most profit could be made.

References

[1] M. Ye, J. Zyren, and J. Shore, 'Forecasting short-run crude oil price using high- and low-inventory variables', Energy Policy, vol. 34, no. 17, pp. 2736–2743, Nov. 2006.
[2] H. Bu, 'Effect of inventory announcements on crude oil price volatility', Energy Economics, vol. 46, pp. 485–494, Nov. 2014.
[3] Michael Ye, John Zyren, Joanne Shore, Forecasting short-run crude oil price using high- and low-inventory variables, Energy Policy, Volume 34, Issue 17, 2006, Pages 2736-2743, ISSN 0301-4215, https://www.sciencedirect.com/science/article/pii/S0301421505001023
[4] J. Baruník and B. Malinská, 'Forecasting the term structure of crude oil futures prices with neural networks', Applied Energy, vol. 164, pp. 366–379, Feb. 2016.
[5] H. Pan, I. Haidar, and S. Kulkarni, 'Daily prediction of short-term trends of crude oil prices using neural networks exploiting multimarket dynamics', Front. Comput. Sci. China, vol. 3, no. 2, pp. 177–191, Jun. 2009.
[6] A. Ghaffari and S. Zare, 'A novel algorithm for prediction of crude oil price variation based on soft computing', Energy Economics, vol. 31, no. 4, pp. 531–536, Jul. 2009.
[7] H. Ghoddusi, G. G. Creamer, and N. Rafizadeh, 'Machine learning in energy economics and finance: A review', Energy Economics, vol. 81, pp. 709–727, Jun. 2019.
[8] Sundria, S. and Smith, G. (2019) Oil Jumps to Four-Week High on Report of Falling U.S. Stockpiles - Bloomberg. Available at: https://www.bloomberg.com/news/articles/2019-06-25/oil-jumps-on-tighter-u-s-supplies-ongoing-middle-east-jitters (Accessed: 26 June 2019).
[9] Palmer, B. (2019) EIA vs. API: Comparing Crude Oil Inventory Reports, Investopedia. Available at: https://www.investopedia.com/articles/investing-strategy/090816/eia-vs-api-comparing-crude-inventories-announcements.asp (Accessed: 26 June 2019).
[10] Field, Discovering statistics using R. Sage publications, 2009.
[11] B. Lantz, Machine Learning with R;Expert techniques for predictive modeling, 3rd ed. Sage publications, 2019.
[12] G. Pierobon, "A comprehensive Machine Learning workflow with multiple modelling using caret and caretEnsemble in R," Medium, 31-Oct-2018. [Online]. Available: https://towardsdatascience.com/a-comprehensive-machine-learning-workflow-with-multiple-modelling-using-caret-and-caretensemble-in-fcbf6d80b5f2. [Accessed: 26-Jul-2019].
[13] [1]M. Kramer, 'Short (or Short Position) Definition', Investopedia. [Online]. Available: https://www.investopedia.com/terms/s/short.asp. [Accessed: 11-Aug-2019]. 12-May-2019.