Predicting Oil Generation Using Supervised Learning Algorithms

Student Mubarak Babslawal
Course Statistical Learning
Programme Data Science in Economics
Github Repo https://github.com/Mubarakbabs/SL_final_project/

Abstract

The objective of this analysis is to accurately predict future oil production from a set of oil wells using multiple machine learning models. Accurate predictions can assist oil producers in forecasting production to meet future demand, while also identifying key factors that influence production. Three models—Linear Regression, Decision Trees, and Random Forest—were employed to fit the dataset, which consists of 6925 observations across 14 variables. The dataset underwent preprocessing, including imputing missing values and feature engineering. Key findings reveal that time-related variables significantly affect prediction accuracy, and Random Forest outperformed the other models with an adjusted R-squared of 79.7%. The study also highlights the importance of feature engineering in improving model performance.

1 Statement of the Problem

The aim of the analysis is to predict oil production for a group of oil wells using various analysis. These predictions, if accurate can help an oil producing organization to project how much oil they will be able to produce to meet demand over time. Also, from a technical perspective, identifying the factors that most impact production can help the production engineers identify levers to modify to achieve a desired level of production. To achieve this, the data set was fitted on three models, and then the performance across them was compared

1. Linear Regression
2. Decision Trees
3. Random Forest

2 Description of Dataset

The dataset consists of 6925 observations across 14 columns, 13 features and one target: oil production.

```
 'data.frame':  6925 obs. of  14 variables:
 $ PRODUCTION.DATE              : chr  "07/04/2014 00:00" "08/04/2014 00:00" "09/04/20
...
 $ Field.Name                   : chr  "DSEAT" "DSEAT" "DSEAT" "DSEAT" ...
 $ WELL_BORE_CODE               : chr  "DSEAT-001-F-1 C" "DSEAT-001-F-1 C" "DSEAT-001-
 $ N_WELL_BORE_CODE             : int  105 105 105 105 105 105 105 105 105 105 ...
 $ WellBore.Name                : chr  "001-F-1 C" "001-F-1 C" "001-F-1 C" "001-F-1 C"
 $ FLOW_KIND                    : chr  "production" "production" "production" "product
 $ WELL_TYPE                    : chr  "OP" "OP" "OP" "OP" ...
 $ Downhole.Pressure..PSI.      : num  0 0 0 0 4500 ...
 $ Downhole.Temperature..Kelvin.: num  273 273 273 273 370 ...
 $ Average.Tubing.Pressure       : num  0 0 0 0 4021 ...
 $ Annulus.Pressure..PSI.        : num  0 0 0 0 0 0 0 0 0 0 ...
 $ AVG.WHP..PSI.                 : num  0 0 0 0 480 ...
 $ Choke.Size                    : num  0 0 0 0 33.1 ...
 $ Oil.Production..stb.day.      : num  0 0 0 0 0 0 0 0 0 0 ...
```

Six of the features are numeric, while five are character vectors. From among the character vectors, three are simply different ways of naming the oil wells while the flow kind variable and well type have the same values for all the observations. The production date should actually be a date, rather than character. So we're left with six numeric features, one time dimension and one character feature.

The dataset has no missing values except for choke size, which had six missing values. This was handled by imputing the missing values with a 7-day rolling average of the variable.

2.1 Time Range

The dataset consists of data from 2008-02-12 to 2015-06-30

2.2 Well Bore Code

The well bore code variable gives us the names of the oil wells. The oil wells were introduced on the following dates:

```
 well_bore_code       min_date
<chr>                <date>

DSEAT-001-F-1 C          2014-04-07
DSEAT-001-F-11 H         2013-07-08
DSEAT-001-F-12 H         2008-02-12
DSEAT-001-F-14 H         2008-02-12
DSEAT-001-F-15 D         2014-01-12
```

This means that the number of observations available for each oil well is different, which may affect the accuracy of predictions

3 Key Findings

The Key Findings from the analysis showed that:

1. Most of the variables, especially the time dimension played a significant role in predicting oil production
2. Random forest does the best job in forecasting future values
3. Feature engineering can have a huge impact on model performance

The reasoning behind each key finding is presented in the section below

4 Analysis of Key Findings

This section describes the analysis performed to arrive at the key findings above. The analysis is organized under three sections

1. Significance of variables
2. Comparison of models
3. Importance of Feature Engineering

4.1 Significance of variables

The R markdown script attached to this article goes through the various procedures that were attempted to examine the significance of various variables in the model. The final summary of the linear regression model is presented below showing each variable's significance in the model

```
    Call:
lm(formula = y ~ . + I(days_from_origin^2), data = train_set)

Residuals:
     Min       1Q   Median       3Q      Max
-26403.1   -524.5    -24.8    431.5  22327.2

Coefficients: (1 not defined because of singularities)
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      3.220e+04  5.187e+03   6.208 5.77e-10 ***
production_date                 -3.740e+00  3.046e-01 -12.279  < 2e-16 ***
well_bore_codeDSEAT-001-F-11 H   7.008e+02  2.798e+02   2.504   0.0123 *
well_bore_codeDSEAT-001-F-12 H  -1.501e+03  6.049e+02  -2.482   0.0131 *
well_bore_codeDSEAT-001-F-14 H  -2.775e+03  5.860e+02  -4.736 2.24e-06 ***
well_bore_codeDSEAT-001-F-15 D  -1.835e+03  2.975e+02  -6.167 7.45e-10 ***
downward_pressure_psi           -1.864e+00  2.493e-01  -7.477 8.79e-14 ***
downhole_temp_kelvin             9.220e+01  7.118e+00  12.952  < 2e-16 ***
avg_tubing_pressure             -2.862e-01  1.850e-01  -1.547   0.1220
annulus_pressure_psi             3.824e-01  3.988e-01   0.959   0.3376
avg_whp_psi                      4.947e+00  3.475e-01  14.236  < 2e-16 ***
choke_size                      -8.223e+01  5.937e+00 -13.850  < 2e-16 ***
lag_y                            7.237e-01  8.381e-03  86.353  < 2e-16 ***
days_from_origin                        NA         NA      NA       NA
I(days_from_origin^2)            6.034e-04  1.072e-04   5.628 1.92e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2757 on 5521 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.9133,    Adjusted R-squared:  0.9131
F-statistic:  4473 on 13 and 5521 DF,  p-value: < 2.2e-16
```

Most of the variables are significant in the prediction of oil production. However, there are only two variables that are not significant at all: average tubing pressure and annulus pressure. This can be explained by the level of correlation. While the level of correlation is not high enough for R to ignore the variables entirely, it can be seen that average tubing pressure correlates strongly with downward pressure. Hence, downward pressure has sufficiently explained the variation that could have been explained by average tubing pressure

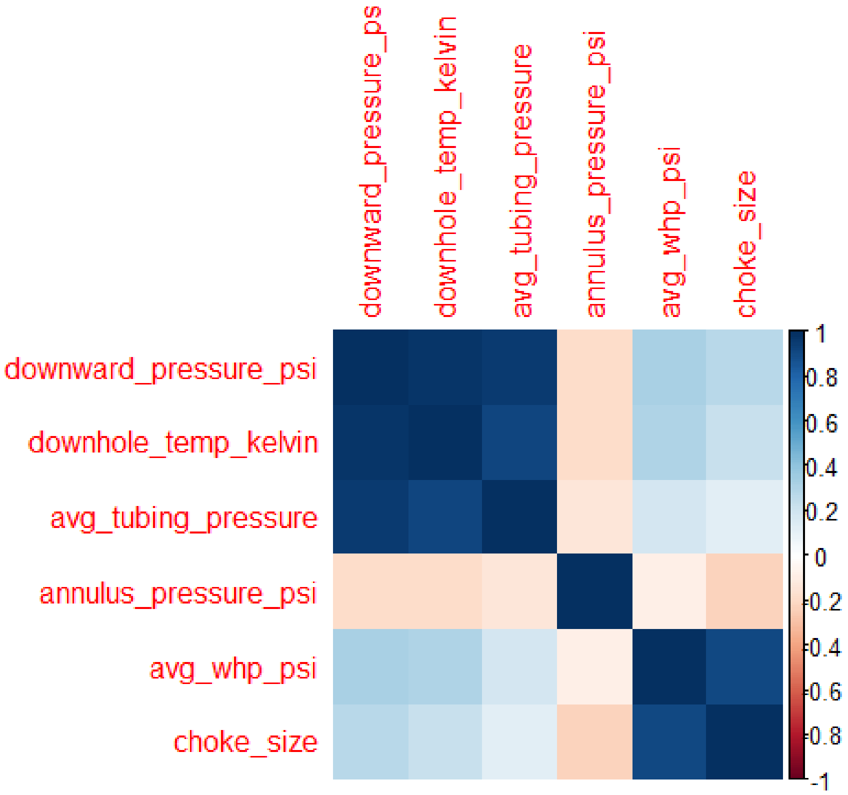

Figure 1: Correlation

The `days_from_origin` variable isn't considered in the linear model because it carries the same information as the production date.

4.2 Comparison of models

The primary metric for comparing the models is the adjusted r-squared on the test set. A model's performance must be tested on data that it hasn't seen in training to evaluate its efficacy on a new data set. The r-squared works well as an evaluation tool in a regression context because it measures how far each prediction is from the actual values, and squares the differences, hence penalizing predictions farther from expectations more heavily than those closer to the actual values. Below are the R-squared values from each of the models:

Model	**Adjusted R squared**
Linear Regression	75.7%
Decision Tree	66.5%
Random Forest	79.7%

Table 1: Performance of the three models

The performance of each model can be explained by looking at a chart comparing the actuals with the predicted variables

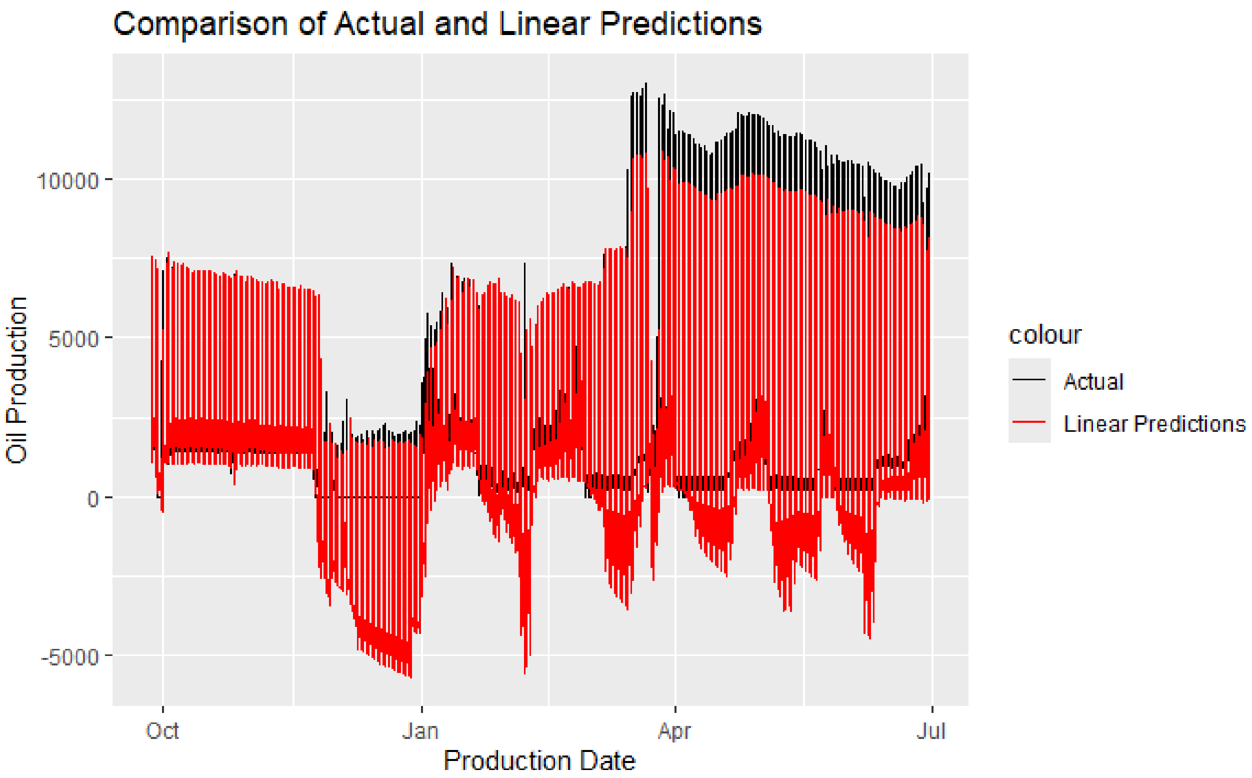

Figure 2: Linear Predictions vs Actual

Linear Predictions Performance The first few months of predictions in the dataset seemed to have way higher predictions than the actual. The linear prediction generally followed the pattern of the dataset, and many times predicted far higher than the actual production. This changed in the last few months, where the actuals were all higher than the predictions. This phenomenon is also present in the other models. This is likely because the oil production was much higher in those months than any other months. Since the linear regression generally assumes that a certain trend wil continue throughout the dataset, it doesn't increase its predictions sufficiently to match the higher oil productions when this trend changed.

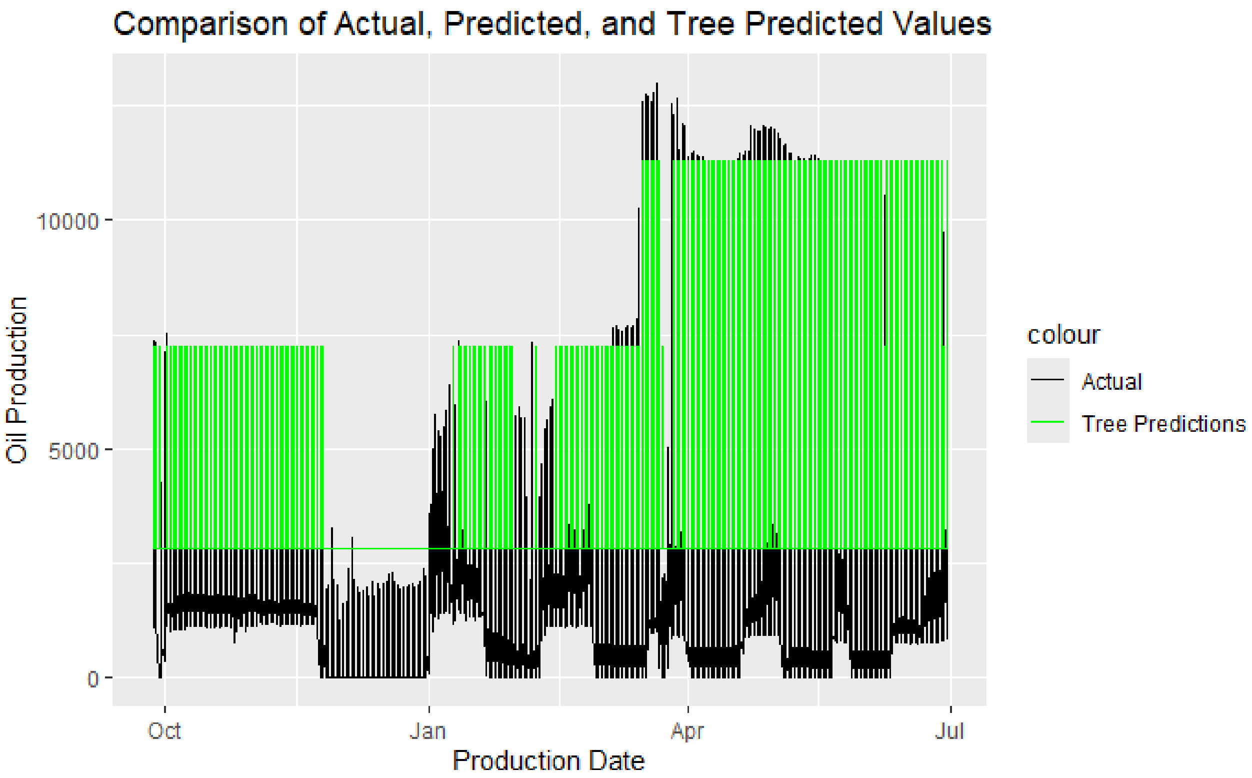

Figure 3: Decision tree performance

Decision tree Performance The single decision tree has decisively the worst performance on the dataset among the models analyzed and Figure 3 shows why. Regression decision trees don't predict individual values, rather they predict the values for a range of inputs, where the values are expected to be similar. Hence the decision tree doesn't sufficiently react to the turbulence of the actual dataset. This means that it is more likely to hold on to incorrect values for a larger amount of values, and so its residual sum of squares is much higher.

Random Forest Performance One way of handling the shortcomings of a decision tree is by using an ensemble of decision trees: i.e. a random forest model. The random forest performs best among the three models evaluated:

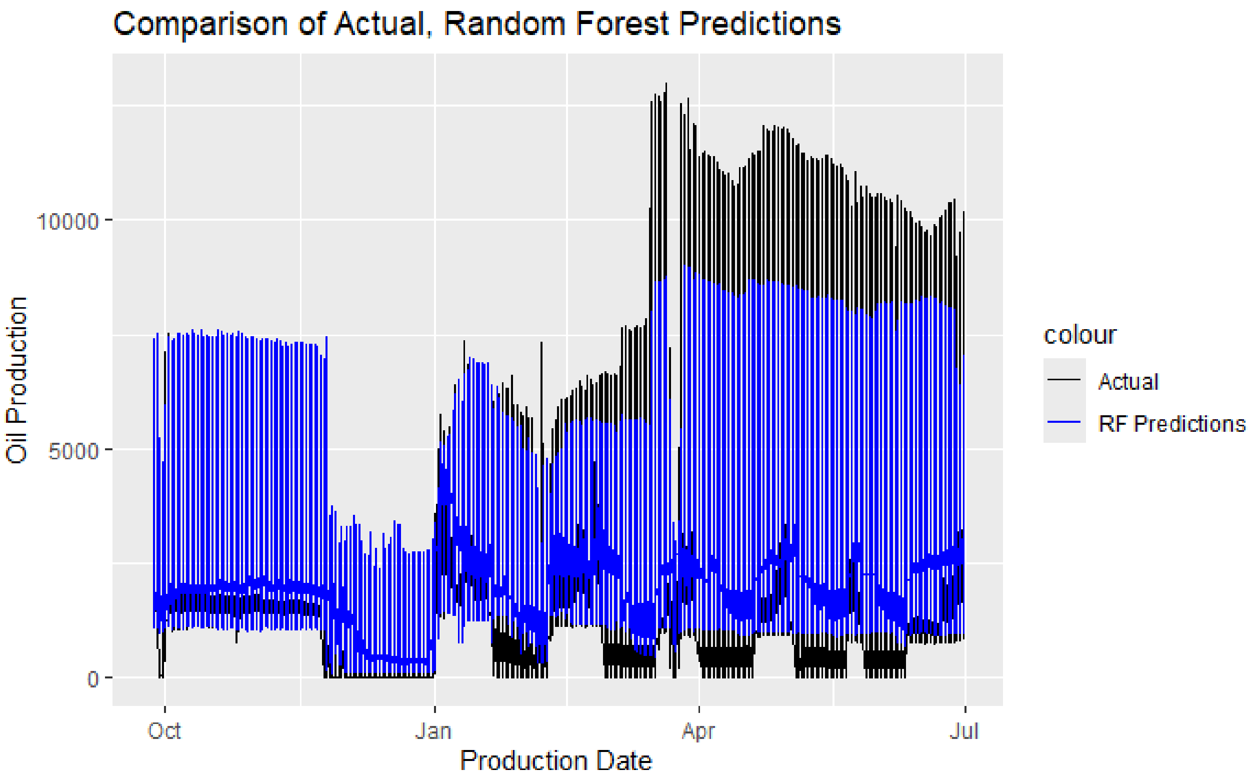

Figure 4: Random forest performance

As seen above, the random forest does a better job of reacting to the changes in the values than a single decision tree. Thus, in areas where the values of the actual oil production change only slightly, the random forest is able to model the data almost perfectly. However, in the areas where there is a significant change in the actual values, just like the linear regression model, the random forest doesn't react sufficiently to meet up with the new actual target values.

MSE and MAE

However, the R-squared is not the only way of measuring the performance of models on regression tasks. One metric that can also be used to compare performance of linear regression models is the Mean Absolute Error and Mean Squared Error

Model	MAE	MSE
Linear Regression	1032.263	2223698
Decision Tree	1528.196	3070756
Random Forest	1052.979	2147167

Table 2: Performance of the three models

While the random forest has the best R squared value as well as the best squared error, the mean absolute error is lowest in the linear regression model. Remember that the mean squared error squares the values before taking an average, so larger differences are penalized even more. The extra penalization due to larger differences affects the linear regression more than the random forest. As seen in 4, random forest seems to be more conservative and tends to maintain a similar prediction over a longer period. Hence it is less likely to have preditions with very high deviations. The lower prevalence of extremely large deviations in the random forest model explains why it has a lower MSE even though its MAE is higher than in the linear model.

4.3 Importance of Feature Engineering

Alongside the type of machine learning model used, the way data is preprocessed can have a huge impact on the outcome of the model-building process. A good example of this in the analysis was the introduction of the lag_y variable that uses the previous value of the target variables to predict the next. This transformation alone was responsible for taking the r-squared on the training data set from 79% to 91%

5 Conclusions

From the analysis in the previous section, we're able to arrive at the conclusion that a random forest model performs best on the regression task on average. However, for accuracy of individual predictions, it might be worth considering linear regression since it has the lowest MAE. The report also showcases the importance of data preprocessing, and feature engineering. It can be very useful to take a step back and find out how the same data can be expressed in a different format that will improve the training of the machine learning model.

6 Appendix

You can access the r markdown file containing the step-by-step analysis here

The dataset is linked here