# A comparison of Linear Regression and Random Forest Regression for a Combined Cycle Power Plant Data Set

Quy Vu and Daniel Devine

## Description and motivation for the problem:

- Examine the relationship of Temperature (T), Ambient Pressure (AP), Relative Humidity (RH) and Exhaust Vacuum (V) to predict the net hourly electrical energy output (PE) of the plant.
- Use Random Forest and Linear Regression to predict the response variable PE, then compare the performance of the two algorithms.

## Initial Analysis including basic statistics:

- Dataset: Combined Cycle Power Plant from UCI
- The dataset has 4 continuous predictors and 1 response variable.
- No duplicate or redundant indicators were identified.
- The correlation coefficient matrix shows the potential linear relationship between the indicators *(Table 1)*
- Normalisation is considered as the variables are not on the same scale.
- The ambient temperature (AT) and the vacuum variable (V) have a highest chance of a linear relationship *(Table 3)*
- Histograms of the variables shows a decent normal distribution shape.
- Visual inspection shows insignificant outliers

*Table 1: Correlation Coefficient Matrix*

| | AT | V | AP | RH | PE |
|---|---|---|---|---|---|
| AT | 1 | | | | |
| V | 0.8441 | 1 | | | |
| AP | -0.5075 | -0.4135 | 1 | | |
| RH | -0.5425 | -0.3122 | 0.0996 | 1 | |
| PE | -0.9481 | -0.8698 | 0.5184 | 0.3898 | 1 |

1

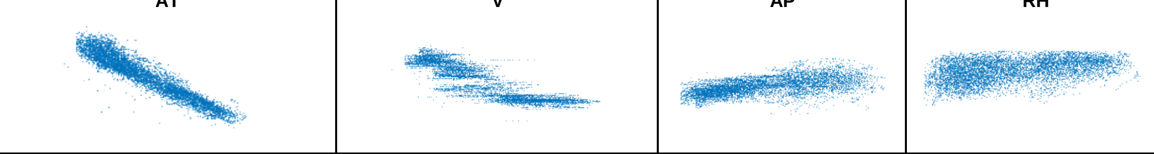
0

-1

*Table 2: Summary of Basic Statistics*

| | AT | V | AP | RH | PE |
|---|---|---|---|---|---|
| mean | 1.81 | 25.36 | 992.89 | 25.56 | 420.26 |
| med | 20.35 | 52.08 | 1012.94 | 74.98 | 451.55 |
| max | 37.11 | 81.56 | 1033.30 | 100.16 | 495.76 |
| min | 1.81 | 25.36 | 992.89 | 25.56 | 420.26 |
| range | 35.30 | 56.20 | 40.41 | 74.60 | 75.50 |
| stdev | 7.45 | 12.71 | 5.94 | 14.60 | 17.07 |
| skew | -0.14 | 0.20 | 0.27 | -0.43 | 0.31 |

*Table 3: Strong linear patterns*




## Summary of the two machine learning models with their pros and cons:

### LINEAR REGESSION ("LR")

***General:***
- Parametric model
- Built on several assumptions about the data *(see "Cons" section)*
- Widely used and ranked as one of the most important tools used in the biological, neurological and social sciences [4].

***Pros:***
- Short run time, which allows premature results to be quickly produced.
- Easy to interpret as the coefficients for the variables give an explanation of their relative importance [1].
- Performs well in high-dimensional data provided there are many observations. [4]

***Cons:***
- Unable to handle large numbers of features with a relatively small numbers of observations [7].
- Produces modest results if the following assumptions are violated: data is normally distributed, heteroskedasticity, linear relationship or independence in data.
- Limited to predicting numerical output.
- Vulnerable to outliers when not accompanied with robust methods and regularisation.
- Requires feature engineering which is problematic in high dimensional data.

### RANDOM FOREST ("RF")

***General:***
- Non-parametric model
- Builds an ensemble of decision trees, each perform on a subset of features, takes an average value of the predictions - gives a numerical result.
- Can choose an attribute as a random root node or pick attribute depending on information gain.
- Considered state-of-the-art

***Pros***
- Allows the maintenance of high dimensionality in the dataset.
- Well regarded in terms of accuracy - particularly important in computer vision.
- Captures the non-linear nature of the data and assimilates it [2].
- Can predict both numerical and non-numerical output.
- Address overfitting problems with a reasonable number of trees.
- Handles noise and missing data well.

***Cons:***
- Interpretability is traded for accuracy.
- RF cannot train outside values in the range of the training data. As the predictions in random forests are provided by taking a average of the results of several trees. [4]
- Needs a large number of trees to extend its full predicting potential. However as the model increases in complexity so too does its computational cost.

## Hypotheses Statement

- The effect of ambient temperature (AT) on output (PE) is a powerful predicting factor with a correlation score of -0.95.
- For LR, the best accuracy is obtained when all features are used [1].
- Both algorithms can produce solid predictions, given a clear and consistent linear trend in the data.
- RF is renowned as a state-of-the-art algorithm which edges out LR.

## Description of Choice of Training and Evaluation Methodology

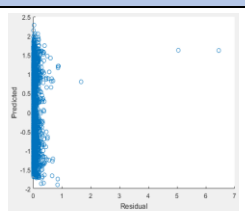
- Dataset is comprised of 9815 data points. Hold out ratio is 0.2
- Data were normalised to run the LR. In both cases the data points were initially randomised.
- Outliers are negligible so robust methods were not necessary.
- To find the optimal RF model, we train the model with the number of trees from 1 to 100 and the number of variables selected as sample from 1 to 3, which is 300 times in total.
- Models will be compared using R-squared statistics, RMSE and run-time.

## Choice of Parameters and Experimental Results:

### LINEAR REGESSION

***Parameters:***
- P-Value for all attributes was less than 0.05 - selected all values.

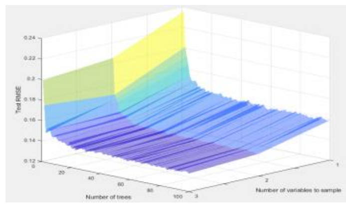
***Main Experimental Results:***
- The linear model is as follows:
  PE ~ 4.819 - 0.86AT - 0.17V + 0.02 AP - 0.13RH
- The large value of intercept and the residual plot means there is still uncaptured pattern and a few outliers.
- The model performed slightly better in training data but still provided good prediction to unseen data *(Table 4 and 5)*




### RANDOM FOREST

***Parameters:***
- The number of trees are in range of 1 to 100
- The number of variables to sample is in range of 1 to 3

***Main Experimental Results:***
- The model performed slightly better in training data,
but still provided good prediction to unseen data. *(Table 5 and 6)*
- The optimal model has 74 trees, taking 3 variables to sample - which is in line with the research conducted by Oshiro et al. [5]




***Further Results of RF and LR:*** *(Table 7)*
- Revising the hold out ratio to 0.3 and 0.1 did not significantly affect the performance of the model.
- Removing either AP, RH or both decreases the performance of the model.

*Table 7: Further results - Test RMSE*

| Model | Feature removal | | | | | Hold out ratio | | |
|---|---|---|---|---|---|---|---|---|
| | N/A | AP | RH | | AP, RH | 0.3 | 0.2 - Original | 0.1 |
| LR | 0.277 | 0.295 | 0.295 | | 0.3 | 0.276 | 0.276 | 0.2756 |
| RF | 0.134 | 0.145 | 0.144 | | 0.16 | 0.137 | 0.133 | 0.1333 |

*Table 4: LR - Main result, predicted (red) vs. true value (blue)*

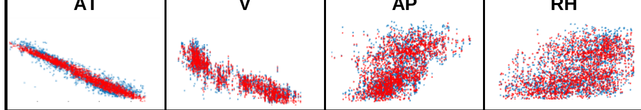
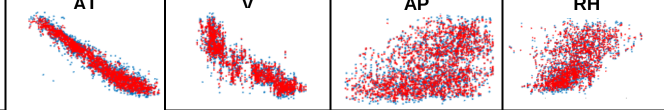



*Table 5: LR and RF performance summary*

| LR | | Criteria | RF | |
|---|---|---|---|---|
| *Train* | *Test* | | *Train* | *Test* |
| 0.269 | 0.277 | RMSE | 0.130 | 0.134 |
| 0.927 | 0.722 | R-squared | 0.869 | 0.866 |
| 1.8 | | Run time (s) | 1527 | |

*Table 6: RF - Main result, predicted (red) vs. true value (blue)*




## Analyses and critical evaluation of the results

1. LR performed well comparing to RF. Since the assumptions for LR were more or less satisfied, the full 'Prediction Power' of RF was not fully utilised in this case. The reasons are as follows:
  + Consistent and linear relationships were noticed throughout the data. This indicates that there was not a presence of local patterns, which RF is powerful in identifying [1].
  + Heteroskedascity can be noticed, and outliers were insignificant [6].
  + The dataset has 4 features and all of them were statistically significant. RF produces accurate results when random features are chosen as the root node. The ability of RF to handle random root nodes and splitting nodes is thus not of strong importance in this case.
  + The dataset is comprised of continuous numerical variables only, while RF can deal with other types of variables containing discrete and categorical variables [9].

2. The run time of the RF model with only 1 tree is 7 seconds, which is 3 times more than the time taken to run a linear regression model (1.8 seconds). To find the optimal number of trees, we ran the model 300 times which took 20 minutes. RF is therefore more computationally intensive.

3. Both models showed good fit to the test set, with the RMSE indicators of 0.2772 and 0.1340 for LR and RF respectively. Given the range of the value to be predicted (PE) is from -2 to 2.4, these numbers indicate an error of roughly 1%. This means that the models generalises well to unseen data - which conforms with the initial hypothesis.

4. Modifying the hold out ratio did not significantly affect the models' performance. This is because of the high observations to features ratio (9815 to 4) and the simple relationship in the data. The number of observations is large enough for both models to capture most of the pattern in the dataset.

5. Although the scatter plot and the correlation matrix did not show significant patterns between AP and RH, and the output (PE), removing one of them from the model clearly decreased the performance on the test set, seen through an increase in the test RMSE by 6.5% for LR and 8.1% for RF. However, given the strength of our models' original performance, such error is tolerable, especially when computational cost is a factor.

6. Breiman suggested that the number of predictors taken as sample for RF should be one third for regression models, however the result showed a declining trend of RMSE as the number of predictors taken as sample increases. This may seem like a counter-evidence, but with limited input variables (i.e 4), ommiting any of them could mean noticeable information loss.

7. In terms of interpretability, the high value of intercept and the residual plots of LR clearly pointed out that there are still uncaptured patterns present in the data. This is true because information on the gas and steam turbines which generate power is not included in the dataset [1]. However, we wouldn't be able to derive such conclusion from RF.

## Lessons learned and future work

***Lesson learned:***
- Both models neglect that changes may occur in the levels of interaction between a system of variables [1]. We know that PE is dependent on variables AT,V,AP & RH, but can also be certain that there are other interactions occuring within the system. These complex interactions between variables are something that are not considered in either model. For example the temperature variable (AT) can affect the interaction between other variables differently in summer and winter. Kaya et al. outlines that in this case models with localised observations can prove to be more beneficial.

- For simple datasets, a simple model will produce accurate results whilst also being computationally efficient, thus is preferred. Random Forest model goes against Ockams Razor theory and incorporates as much as possible into the system.

- On the other hand for a dataset with high dimensionality, implementing a Linear Regression model would require feature engineering and other data preprocessing steps such as data normalisation, data transformation, handling of missing values and outliers. In contrast these initial steps may not be neccessarily required in a Random Forest model. It is therefore important to understand the dataset and select the most reasonable analytical approach. The lesson to take from this is that "there is no free lunch".

- To save time and computational resources, it is best to first implement a simple model on the most significant predicting variables. If the result is positive, it may not be necessary to fit a complex, expensive and hard-to-interpret model since the result may not be significantly improved.

***Future work***
- We acknowledge from the research work carried out that feeding additional observations into a Random Forest regression model improves performance. This is not the case with linear regression - i.e. the model does not benefit from the addition of more data.

- Generally with more training samples, there is more knowledge for RF to learn and so the model can subsequently capture the non-linearity of the structural data. Also, the model will improve with more appropriate features which can create a higher probability of choosing an optimum splitting feature [2].

- Taking the above into account we can say that in terms of future work the Random Forest method will be superior.

***References***


[1] Heysem Kaya, Pınar Tüfekci , Sadık Fikret Gürgen. Local and Global Learning Methods for Predicting Power of a Combined Gas & Steam Turbine, Proceedings of the International Conference on Emerging Trends in Computer and Electronics Engineering (2012) ICETCEE 2012, pp. 13-18
[2] Li, H., Leung, KS., Wong, MH. et al. BMC Bioinformatics (2014) 15: 291.
[3] Leo Breiman, 'Random Forests', Machine Learning 45, no. 1 (2001) 5-32.
[4] Paul Smith, Siva Ganesh, Ping Liu. A Comparison of Random Forest Regression and Multiple Linear Regression for Prediction in Neuroscience. Journal of neuroscience methods (2013). 220:10.1016
[5] Oshiro, T.M., Perez, P.S. and Baranauskas, J.A., 2012, July. How many trees in a random forest?. In MLDM (pp. 154-168).
[6] Alice Zheng. Evaluating Machine Learning Models. O'Reilly.
[7] James N. Morgan, John A. Sonquist. Problems in the Analysis of Survey Data, and a Proposal. Journal of the American Statistical Association (1963). Vol. 58 , Iss. 302.
[8] Ulrike Grömping. Variable Importance Assessment in Regression: Linear Regression versus Random Forest. The American Statistician 63, no. 4 (2009): 308-19.
[9] Ned Horning. Remote Sensing for Ecology and Conservation: A Handbook of Techniques.