Comparison of Multiple Linear Regression and Random Forest Regression applied to the Absenteeism at Work Dataset

Hisho Rajanathan and Wai kit Yeong

Description and Motivation of the Problem

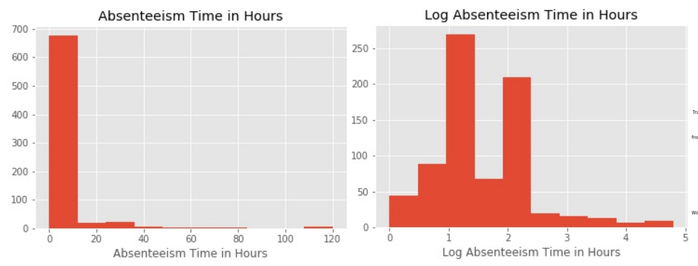
- Compare two popular machine learning algorithms, Multiple Linear Regression and Random Forest in a regression problem, predicting the number of hours of absenteeism at work based on 20 predictors
- Motivation for selecting dataset: High annual costs associated with absenteeism totaling 84 billion USD; other indirect costs include reduced productivity, safety issues, and poor morale among employees [1]
- Aim is to understand the factors of absenteeism and to utilize machine learning techniques to predict estimated hours of absence for an employee given a set of perimeters
- Results compared with Trivedi, Harshit (2018)

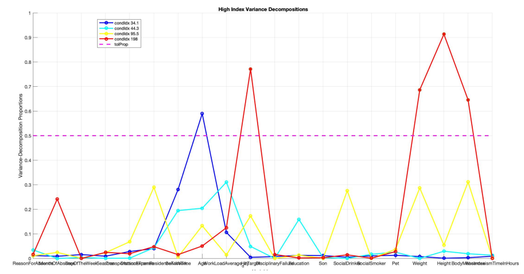
Initial Analysis

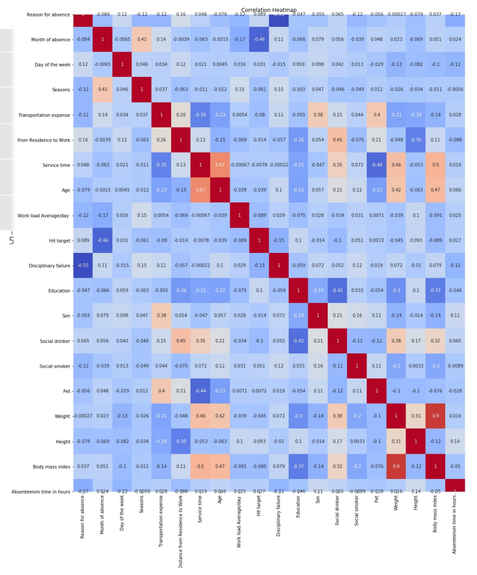
- Dataset: Absenteeism at Work from UCI Repository [2]
- The dataset has 740 rows of data and 20 features, collected from 36 different employees of a courier company in Brazil from a period between July 2007 to July 2010
- The response variable has a large right skew which can be solved by applying a log(x+1) transformation. A log(x+1) transformation has been applied as there are a number of zero hours of absenteeism in the data. Normalisation has not been applied to the response variable due to this being a regression problem.
- The original dataset has 20 predictors which consists of: 9 Categorical Variables and 11 Numerical Variables and no missing data
- Several predictors were not included in the model due to irrelevance and lack of explanation on what they were (e.g. ID)
- Correlation matrix did not show any predictors to be highly correlated with the response variable, hence collinearity was used. This showed Age, Hit Target, Weight, Height & Body Mass Index had a high collinearity so were removed from the dataset
- The mean and standard deviation of absenteeism time in hours was calculated for a sample of continuous parameters

Numerical Variables	Mean	Std.	Max	Min	Skew
Transportation expense	221.5142	66.961	388	118	0.3939
Distance from Residence to Work	29.6242	14.8437	52	5	0.3137
Service time	12.5495	4.3931	29	1	-0.0017
Age	36	6	58	27	0.696
Work load Average/day	271490	39138	378884	205917	0.9575
Hit target	94.5862	3.7869	100	81	-1.2554
Son	1	1	4	0	1.0863
Pet	1	1	8	0	2.6934
Weight	78.9837	12.867	108	56	0.0189
Height	172.1167	6.0453	196	163	2.5572
Body mass index	26.6581	4.2742	38	19	0.3046
Absenteeism time in hours	6.9525	13.3508	120	0	5.7012

Categorical Attributes	Count	Unique	Top	Frequency
Reason for absence	740	28	Medical Consultation	149
Month of absence	740	12	March	87
Day of the week	740	5	Monday	161
Seasons	740	4	Winter	195
Disciplinary failure	740	2	No	700
Education	740	4	High School	611
Social drinker	740	2	Yes	420
Social smoker	740	2	No	686







Description of the two machine learning models and their advantages and disadvantages

Multiple Linear Regression (MLR)

- MLR attempts to model the relationship between variables by fitting a linear equation to the data
- The model is tasked to predict a dependent variable given independent variables via ordinary least squares
- The OLS procedure seeks to minimize the sum of the squared residuals. This means that given a regression line through the data we calculate the distance from each data point to the regression line, square it, and sum all of the squared errors together. This is the quantity that ordinary least squares seeks to minimize.[8]
- The equation represents a Multiple Linear Regression where y is the dependent variable and $x_1, x_2, \dots, x_n$ are independent variables : $y = a_1x_1 + a_2x_2 + a_3x_3 + \cdots + a_nx_n$
- **Pros**: Simple and quick to implement, understandable, results easy to interpret, computationally cheap
- **Cons**: This method assumes the linearity between variables[9], requirement to choose proper variables [10]

Random Forest (RF)

- Random forests consists of a number of decision trees with an ensemble method applied
- RF are random due to i) random subset of the observations, and the split within each tree based on ii) the random subset of variables
- Trees are unstable and hence the randomness allows the individual tree's predictions to be offset by aggregating a larger number of different trees and averaged (for regression) and majority voted (for classification)

Pros: Lower MSE due to induced randomness and averaged prediction, resistant to overfitting and outliers, parallelisable
Cons: Computationally expensive, likely to perform poorly if data contains large number of variables but fraction of relevant variables is small [7]

Hypothesis statement

- We expected that both algorithms will not perform well due to the disconnected nature of the dataset (i.e. lack of relevance of predictors to the response variable, high number of outliers and lack of relationship between the predictors)
- This was supplemented by Harshit [1] who achieved an accuracy of 34% initially when performing a classification study
- Initial thoughts are that random forest regression model will perform better due to results from literature [6], however due to the nature of the dataset, a simpler model (MLR) might gave a smaller MSE if ideal predictor variables are selected
- Concerns on the effect of categorical data such as number of children, type of illness, seasons, etc. may have on the regression analysis may mean that we might remove them from the predictor set

Description of choice of training and evaluation methodology

- Initial training on 7,740 data points after removing redundant parameters
- A 70:30 split was adopted for the train test split. Due to the dataset not being large, 70% of the data will be used for model training
- A comparison was made between the quadratic and linear fit for linear regression model, to see which was the best suited model for this dataset
- Using hyperparameter optimisation such as grid search, random search, Bayesian optimisation to find optimal values for the Random Forest regression model. Random search is recommended due to similar performance and much less computational cost[4]
- 10 fold cross validation was used to estimate generalisation error, RMSE and prevent overfitting
- As this is a regression problem, we will use Root Mean Squared Error (RMSE) and time taken as evaluation metrics [5]

Parameter choice and experimental results

Random Forest (RF)

- Cross validating ensemble of regression trees using 10 fold cross validation result: The ensemble achieves an MSE of around 157.2 after the first learning cycle
- The cross validated MSE initially starts at 157.2, as more learning cycles are accumulated it increases to around 235 and then increases after a small dip
- Mean Squared Error of the baseline model: 6.9727
- Estimated generalisation error of the ensemble = 242.6956

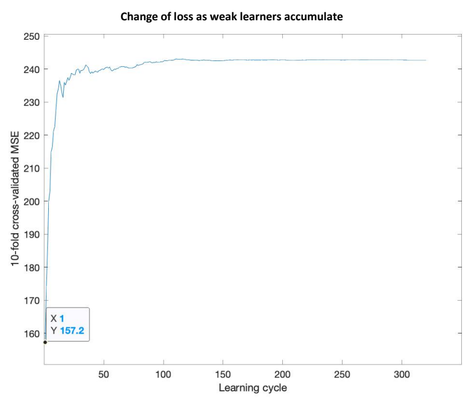
Hyperparameter optimisation using Bayesian optimisation results:

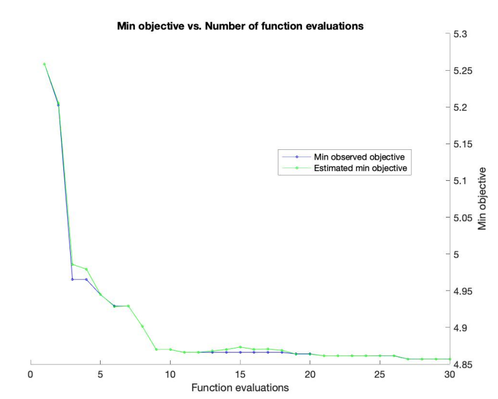
- Optimal Parameter Values: i) Method = LSBoost, ii) Minleafsize = 1, iii)Learning Rate = 0.0061128, iv) learningcycles = 320
- Normalising made no difference to the results
- After the biggest drop from 1 to 3 function evaluations, the minimum objective drops to a low of 4.8571 after 9 function evaluations and decreases marginally after
- RMSE for train set and test set for the fitted RF model are as follows:
  - Train: 2.5385
  - Test: 21.1442

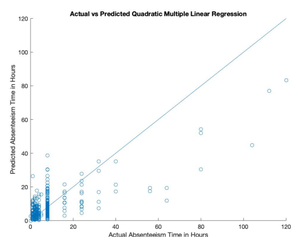
Multiple Linear Regression

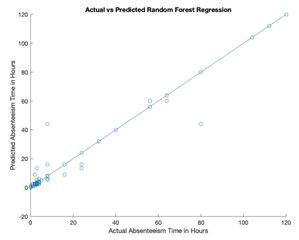
- Multiple linear regression model comparing the suitability of a linear or quadratic fit to the data points.
- Ordinary R Squared statistic for quadratic type is higher than linear type, 60.10% and 15.90% respectively, showing that the quadratic fit was a better model with the higher R Squared statistic
- The residuals on both linear and quadratic types do not fit a normal distribution as they have fatter tails, similar to Poisson distribution
- The cross validated RMSE falls to 0 at kFold = 5 for the quadratic type, where as the linear does not. This shows that the quadratic fit is a better model for Multiple Linear Regression
- RMSE for train set and test set for the fitted models are as follows:
  - Linear:
    - Train: 11.3945
    - Test: 18.6287
  - Quadratic
    - Train: 7.8482
    - Test: 14.3130
- Overall the Quadratic type linear regression model performed better than the linear

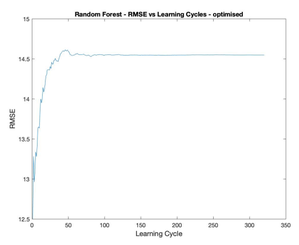
Figures

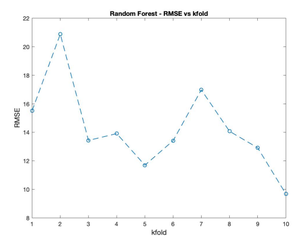


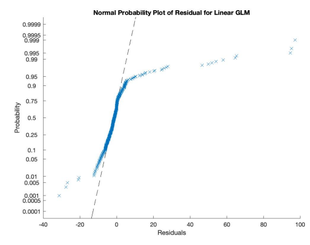


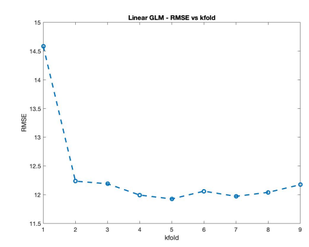


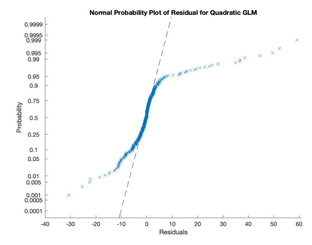


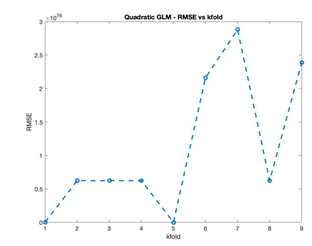












Analysis and Critical Evaluation of Results

- As expected, after running hyperparameter optimisation, the RF model performed better than the MLR model, with an RMSE score of 2.5385 vs an RMSE score of 7.8482 for the quadratic MLR model on the training data.
- However, it required more time to train and implement, with the time taken to run being twice as long as MLR
- This is coherent with [5], which posits that random forests are better able to capture non-linear interaction between the predictors and the target, unlike linear models
- A low learning rate appeared to benefit the performance of the RF model, whereas higher learning rates increased the cross validated MSE scores
- The RF model worked well out of the box, and optimising it only marginally improved its performance – we can deduce that it handled outliers better than MLR
- Varying the features and number of features did not have a significant impact on the RF model. This could be due to the lack of correlation between the variables, even before higher correlated ones were removed from the training set
- The final evaluation metrics could be improved, and more studies are required to select the predictors and tune the hyperparameters. Domain knowledge is critical to run this analysis, as there were a number of variables which did not have sufficient background information (such as disease types), which might be relevant key in predicting hours of absence

Test Set Final Model Results

RMSE

RF	Model	MLR	
		Linear	Quad
2.5385	Train	11.3945	7.8482
21.1442	Test	18.6287	14.3130

Time Taken

RF		Quad MLR
1.44s	Training Time	0.82s
0.39s	Prediction Time	0.04s

- Normalisation was not applied to the dataset as even after applying normalisation to the regression models, the RMSE statistics were unchanged. We believe this is because there is not a high variance with the response variable
- The residuals for the linear regression model, showed that there were fat tails present which implied that the dataset was not normally distributed. Hence, a Poisson distribution was chosen for the linear regression model. This reason also backs up the previous point for not normalizing the dataset
- PCA could be applied to the dataset to better optimise the parameters. PCA was not applied here because there was not a high correlation between the variables
- Using LSBoost as an ensemble method results in negative values in our predictions, which should not be possible as a minimum number of absent hours should be zero. This resulted in a higher MSE and RMSE than expected. A floor was not implemented to cap the minimum values to zero as this can possibly skew the MSE and RMSE results
- This could also explain why the RMSE score for RF was 8.33 times higher for the test set compared to the training set.
- MLR is nearly twice as quick to train the quadratic model compared to using Random Forest Regression. MLR is ten times quicker for prediction time in comparison to Random Forest Regression. Hence Random Forest regression uses more computational power and may not be viable for larger, more complex datasets and where quick computation is required

Lessons Learned and Future Work

- Optimising linear regression leads to more data manipulation. Another method similar to normalisation could be applied to the dataset to improve the metrics
- Multiple linear regression: do comparisons of the model using pure quadratic or polynomial model specification to see if the model improves
- Random Forest Regression tends to require more optimisation of the hyperparameters and grid search and random search could be used instead of Bayesian Search and results compared
- Random Forest Regression: when optimising the model the hyperparameter chosen was LSBoost. If this was changed to bagged, negative values might not be present for the predicted values. This could potentially reduce the difference between the train and test RMSE
- Besides PCA, other techniques for dimensionality reduction could be applied to the dataset to improve metrics and reliability of the models

References

[1] Trivedi, Harshit. "Explaining Absenteeism at Workplace Predicted by a Neural Network." ABCs 2018 - 1st ANU Bio-inspired Computing conference
[2] Absenteeism at Work Dataset, http://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work
[3] Investopedia, "The Causes And Costs of Absenteeism In Workplace," Forbes, updated 26 June 2019 by Jean Folger. [Online]. Available at: https://www.investopedia.com/articles/personal-finance/070513/causes-and-costs-absenteeism.asp
[4] Zheng, Alice, "Evaluating Machine Learning Models: A Beginner's Guide to Key Concepts and Pitfalls", O'Reilly Media, Inc., Sebastopol, CA (2015)
[5] M. S. Acharya, A. Armaan and A. S. Antony, "A Comparison of Regression Models for Prediction of Graduate Admissions",(2019), International Conference on Computational Intelligence in Data Science (ICCIDS), Chennai, India, 2019, pp. 1-5.
[6] N. Monburinon, P. Chertchom, T. Kaewkiriya, S. Rungpheung, S. Buya and P. Boonpou, "Prediction of prices for used car by using regression models", (2018) 5th International Conference on Business and Industrial Research (ICBIR), Bangkok, 2018, pp. 115-119.
[7] Hastie, Trevor, Tibshirani, Robert and Friedman, Jerome. The Elements of Statistical Learning. New York, NY, USA: Springer New York Inc., 2001.
[8] Brownlee, J. (2016). Linear Regression for Machine Learning. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/linear-regression-for-machine-learning/ [Accessed 18 Nov. 2019]
[9] M. Kayri, I. Kayri and M. T. Gencoglu (2017) "The performance comparison of Multiple Linear Regression, Random Forest and Artificial Neural Network by using photovoltaic and atmospheric data," *2017 14th International Conference on Engineering of Modern Electric Systems (EMES)*, Oradea, pp. 1-4.
[10] Marill, Keith. (2004) Advanced Statistics: Linear Regression, Part II: Multiple Linear Regression. Academic emergency medicine : official journal of the Society for Academic Emergency Medicine. 11. 94-102. 10.1197/j.aem.2003.09.005.