***Write Up – Machine Learning Coursework***

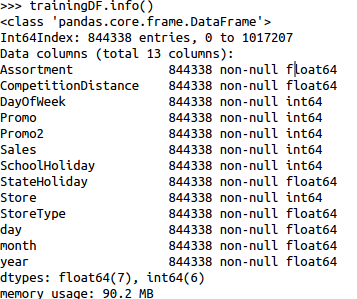
***Brief Description and Motivation:***

Predictive Analytics can be argued as of the most alluring and effective methods of forecasting the future given past observations for all businesses and organisations. To quote Barton and Court ([2012](http://onlinelibrary.wiley.com/doi/10.1111/jbl.12010/full" \l "jbl12010-bib-0001)) [7], “Advanced analytics is likely to become a decisive competitive asset in many industries and a core element in companies' efforts to improve performance.”. Machine style solutions offer far more robust methods for this type of analysis, as they are able to take in account many variables (high dimensionality) with an accuracy that can far exceed the capability of any human, even one with specialist domain knowledge of the particular area.

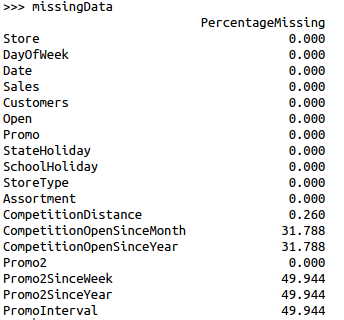
The aim of this analysis is to explore, compare and contrast two Machine Learning algorithms for the purpose of forecasting (regression) the daily sales of 1,115 stores across Europe for the Rossman Drug Stores. The intention is to provide evidence that once particular algorithm is more effective that other methods for the purpose of this task. In conjunction to this there will be a short analysis evaluating if a particular method of determining hyper-parameters can be established.

***Initial Analysis of the Dataset including Basic Statistics:***

* 11 Binary Columns, 3 Numerical Columns, 2 Date Columns, 2 Categorical Columns
* Overview of Dataset



* Missing Data



\*Feature Engineering Ideas\*

1. Features Created

* Mean Sales per Month & Year (\*t-statistic\* → validation)
* Means Sales per Month & Year & Ration to Store (\*t-statistic\* → validation)
* Mean On/Off Promotion & Ratio (\*t-statistic\* → validation)
* Mean On/Off School Holiday & Ratio to Store (\*t-statistic\* → validation)
* Mean On/Off State Holiday & Ratio to Store (\*t-statistic\* → validation)
* Mean On/Off Weekend & Ratio for Store (\*t-statistic\* → validation)
* Log Sales – Already Implemented

2. Relationships Explored

* Each Variables relationship to Predictor (Sales)
* Feature importance?
* Principal Component Analysis (PCA)

***Hypothesis:***

Does a computationally more expensive (Random Forest) Machine Learning algorithm offer a better trade-off in terms of time to optimise and performance than an a less computationally expensive algorithm (Decision Tree)? As a sub-task, does Bayesian Optimisation offer a viable alternative for reducing the the difficultly of determining optimal hyper-parameters. Where optimal parameters will be evaluated based on the output of the cost function.

Bayesian Optimisation was originally developed in the 1970's by Jonas Mockus [1] as a technique for finding global maxima of black-box type functions. It hasn't been until recently with the advent of the Big Data revolution that the need for effective methods of tuning hyper-parameters has been so important, as a result of this Bayesian Optimisation has seen resurgence in popularity ([2],[3],[4],[5],[6]). This method is an alternative to other well know methods such as Grid Search or Random Search, which can be considered brute-force approaches for determining optimal hyper-parameters.

***Description of choice of training and evaluation methods:***

The high level overview of the processes for comparing the two machine learning algorithms can be written as follows:

* *Import dataset*
* *Pre-process dataset*
* *Transform and combine dataset into accessible format*
* *Generate via Feature Engineering a selection of new variables*
* *Define configuration and set-up elements for Bayesian Optimisation (Bounds)*
  + *150 trials per method*
  + *5 Fold Cross Validation*
* *Systematically training, validate and store the mean cost function values*
  + *Train models*
  + *Test models*
* *Return all results, analyse, evaluate and interpret – Graphically and Numerically*
* *Submit test results to Kaggle for comparison*

Preprocessing:

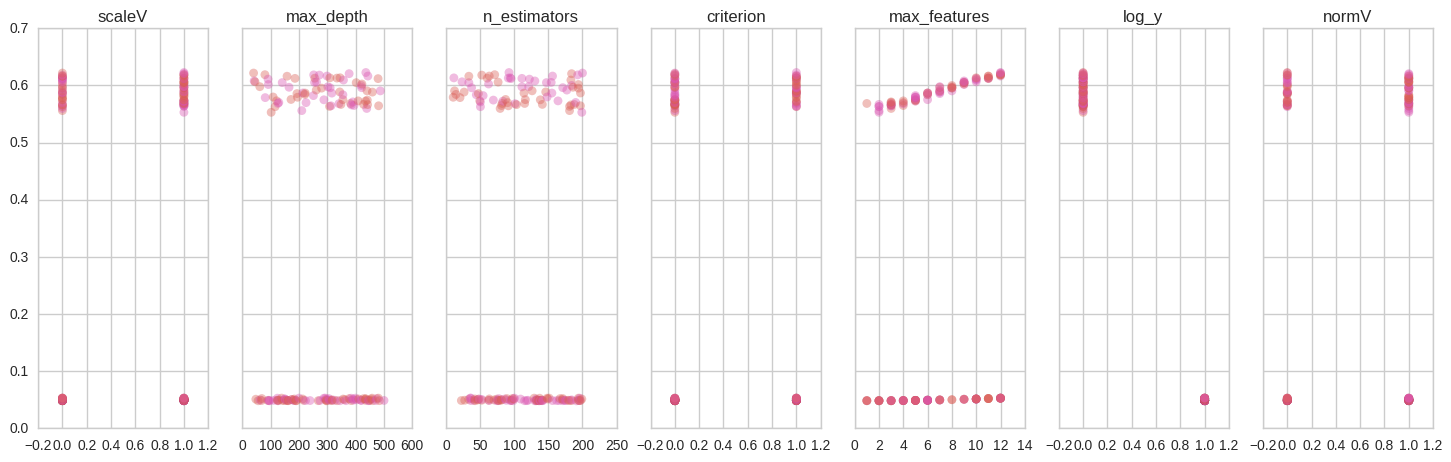
Basic Statistics, Dimensionality reduction considered, feature engineering type analysis, investigation of variables and relationships between them.

Evaluation Phase:

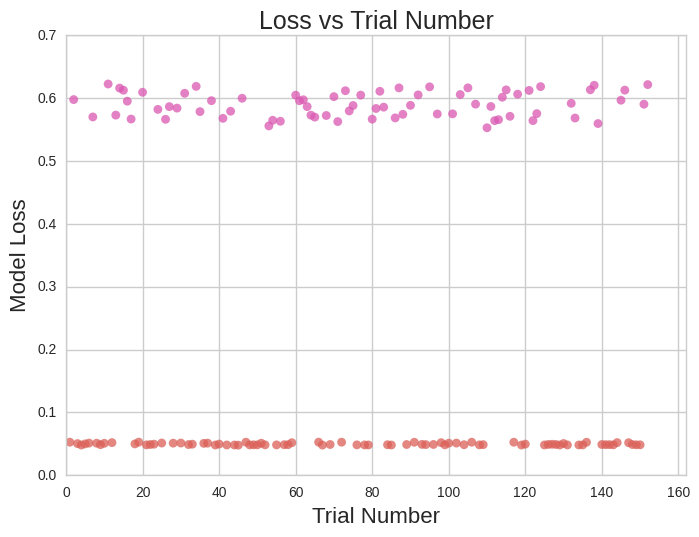
Cross validation is applied for two purposes, firstly to ensure that that the reported performance of an algorithm is as true as possible and also secondly for the aid of model selection. The first point is crucial as a true representation of the average five samples of the training data is better than one. The second point is that the output of each of the trials (mean value of the cost function) will try to be minimised using Bayesian Optimisation, where it is then likely that the global minimum of the trials will most likely be deemed the most effective model. Cross Validation will be restricted to five folds as from initial benchmarking it was found that the time taken to train, predict and evaluate one model both algorithms can range between 40-80 seconds. Therefore it has been decided to use a number of folds that is not to great that the time for the entire processes is not too excessive.

***Analysis and Critical evaluation of results:***

*Parameter by Cost Function:*



Trail Number by Cost Function:



***Lessons learned and future work:***

Lessons Learned:

Future Work:

***References:***

[1] - Jonas Mockus: On Bayesian Methods for Seeking the Extremum. Optimization Techniques 1974: 400-404

[2] - Eric Brochu, Vlad M. Cora, Nando de Freitas: A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning. CoRR abs/1012.2599 (2010)

[3] - Daniel J. Lizotte, Tao Wang, Michael H. Bowling, Dale Schuurmans: Automatic Gait Optimization with Gaussian Process Regression. IJCAI 2007: 944–949

[4] - Frank Hutter, Holger Hoos, and Kevin Leyton-Brown (2011). [Sequential model-based optimization for general algorithm configuration](http://www.cs.ubc.ca/labs/beta/Projects/SMAC/papers/11-LION5-SMAC.pdf), Learning and Intelligent Optimization

[5] - J. Bergstra, D. Yamins, D. D. Cox (2013). Hyperopt: A Python Library for Optimizing the Hyperparameters of Machine Learning Algorithms. Proc. SciPy 2013.

[6] - Snoek, J., Larochelle, H., Adams, R.P., 2012. Practical Bayesian optimization of machine learning algorithms, in: Advances in Neural Information Processing Systems. pp. 2951–2959.

[7] - Barton, D., and Court, D. 2012. “Making Advanced Analytics Work for You.” Harvard Business Review 90:79–83.