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

***Brief Description and Motivation:***

*Brief description and motivation of the problem (5%)*

***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.

***Description:***

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 data set including basic statistics (10%)*

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

As initial phase of the project, a relevant portion of the time has been allocated to the evaluation of the dataset and the dimensions of variables in the contest we were operating. The preprocessing analysis has been conducted in R, n open source software. Aiming to predicts the sales level, a first hint to our process of variable exploration has been given by the study and consequent representation of Sales Level [£] vs Customer Number [#], represented in the following scatter plot.

|  |  |  |
| --- | --- | --- |
|  |  |  |

Has it can be evaluated from the three representation the global Sales vs Customer representation suffers from the bias of including both the subset of data when the promotion is on or off. Consequently, the partition of the subset into the Sales vs Customer during the period of promotion, or not, is more representative of the effect that the independent dummy variable promotion has on Sales and on Customers. It can also be evaluated that when the promotion is on, less people tend to spend more and vice versa when the promotion is off.

The colour attribution is proportional to the density of the data points represented.

Consequently we kept investigating in the nature of the attribution of the average of Sales amount given the individual stores and their relation with the distance given the unconditional distance from competitors, following summarized in the two scatter plots.

|  |  |
| --- | --- |
|  |  |

With the first scatter plot on the left we want to give a representation that during the period of promotion the average sales are significantly bigger than the average sales during non a promotion period and that given the competitors distance, the sales during the promotional period is non affecting the sales level, while it is when the promotion period is off, from the second scatter plot.



With regards with the store type and assortment, as shown in the nearby bar chart, we detected that the type B has the highest average sales and in this, the assortment C is the highest among all other, unconditionally. Store type B is also the only one that has three different assortments (A, B, C) while all the other types of stores have only C and A, having fairly similar average sale in total and in their respective subset by assortment.

The time series represented in the line plot below is the aggregated average level of Customers and Sales respectively. As yearly trend it can be see that there is a seasonal component affecting the increasing trend of both the sales amount and customers’ number that is increasing toward the end of the year. Furthermore we can see that over years, while the number of customers is declining, probably due to an improvement of general population health, the average sales amount is increasing.

|  |  |
| --- | --- |
|  |  |

*Brief summary of the two ML models with their pros and cons (10%)*

***Machine Learning Algorithm 1: Decision Tree***

Prior to introducing the Bayesian Ridge Regression algorithm it is important to note that this algorithm is a variation of the Ordinary Least Squares [OLS]. The objective of the OLS method is to estimate unknown parameters in a regression model, this is achieved through the minimisation of the distances between the actual data and the model (finding optimal parameters). The assumptions of the OLS method state that; the errors have an expectation (mean) of zero, errors are independent and have equal variances [REFERENCE]. When the errors to not have equal variances, this usually occurs when the correlation between features is large, means that the assumptions of the model are violated.

Bayesian Ridge Regression in contrast to the OLS method tries to find the biased estimators and so no longer needs to follow the assumptions of the OLS method. By making use of the the fact that if there exists high variance in the parameters, the overall variance of the model can be lower. Finally by accepting bias into the model it in turns means that the variance of the predicted variables becomes more stable and a better model in some circumstances.

Dan to Add – Strengths and Weaknesses / Add References and Improve the Clarity of the Description

***Machine Learning Algorithm 2: Random Forest***

…Enrico to Expand… SUMMARY, STRENGTHS AND WEAKNESSES

*Hypothesis statement (5%)*

***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 linear least squares algorithm (Ridge Regression) for the task of forecasting Sales for a entire Companies stores? As a sub-task, does Bayesian Optimisation offer a viable alternative for reducing the the difficultly of determining and tuning the optimal hyper-parameters? Where optimal parameters will be evaluated based on the output of the cost function, which will be discussed in the next section.

*Description of choice of training and evaluation methodology (5%)*

***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 – for the purpose of determining the final performance on the Test Set*

***Evaluation Criteria:***

The evaluation criteria that was used for this analysis was the Mean Absolute Error (MAE). Deciding the most appropriate criteria was the decision between the MAE and the Root Mean Squared Error [RMSE], an evaluation criteria that penalizes more where the errors are higher. As concluded in the paper by Willmott and Matsuura [8] “Our analysis indicates that MAE is the most natural measure of average error magnitude, and that (unlike RMSE) it is an unambiguous measure of average error magnitude.”

***Hyper-parameter Optimisation***

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 offers an alternative to other well know methods method of Parameter Optimisation such as Grid Search or Random Search, which could be argued to be brute-force approaches for determining optimal hyper-parameters.

Dan to Add – Briefly how it Works

***Cross Validation:***

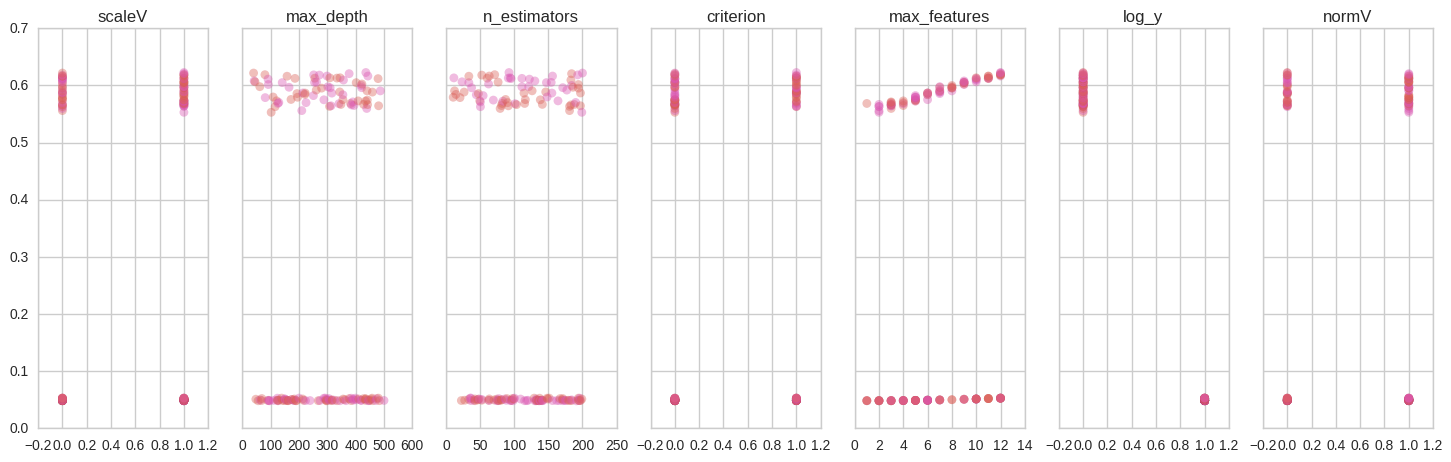
Cross validation is applied for two purposes, firstly to ensure that that the reported performance of an algorithm is as true as possible and also secondly for the aid of model selection. The first point is crucial as a true representation of the average five samples of the training data is better than 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, five, that is not to great that the time for the entire processes is not too excessive.

*Choice of parameters and experimental results (10%)*

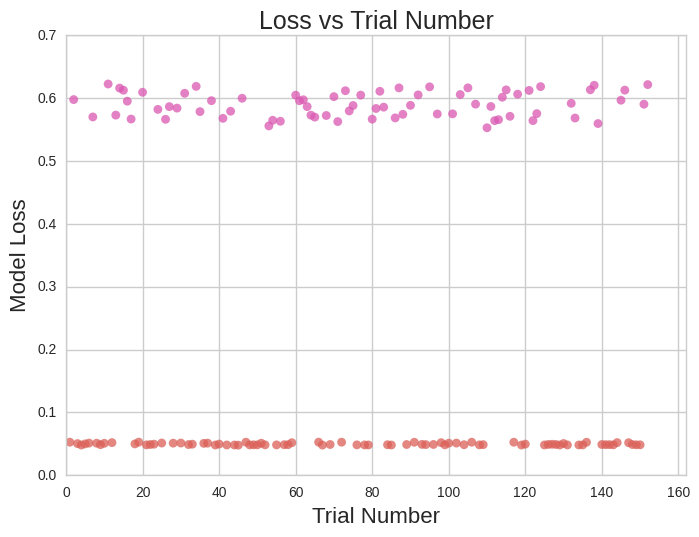
***Define the Bounds of the Hyper-Parameter Search***

Dan to Add

*Parameter by Cost Function:*



Trail Number by Cost Function:



*Analysis and critical evaluation of results (25%)*

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

*Does a computationally more expensive (Random Forest) Machine Learning algorithm offer a better trade-off in terms of time to optimise and performance than linear least squares algorithm (Ridge Regression) for the task of forecasting Sales for a entire Companies stores? ENRICO – Can you advise what you want from the Script to answer this?*

*Does Bayesian Optimisation offer a viable alternative for reducing the the difficultly of determining and tuning the optimal hyper-parameters? DAN – I intend to run Gird Search vs Bayesian Optimisation for the same bounds*

*Lessons learned and future work (5%)*

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

Lessons Learned:

Enrico to Expand <250 words

Future Work:

Time Series Analysis… Enrico to Expand <150 words

***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.

[8] - Willmott, C.J., Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate research 30, 79.