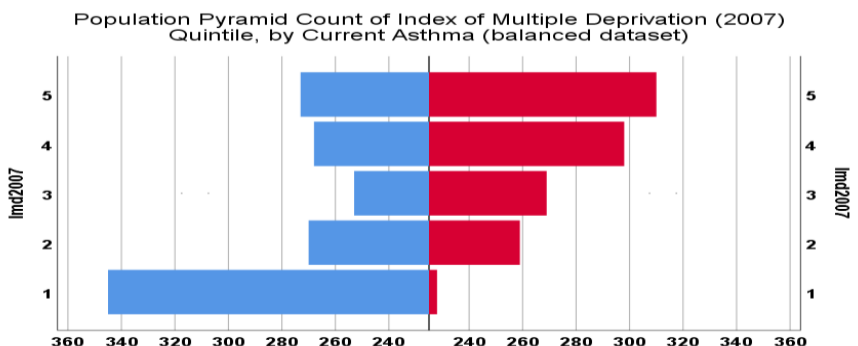
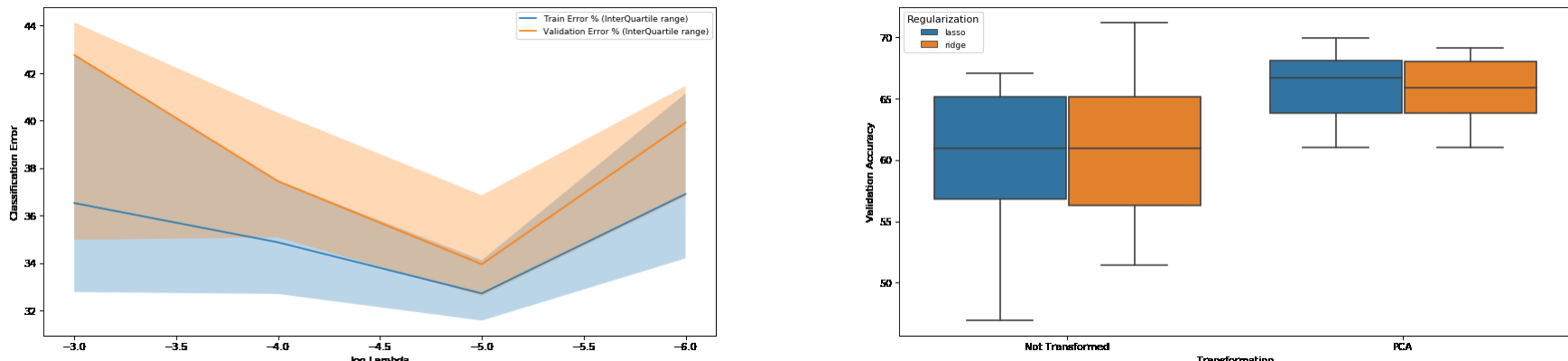
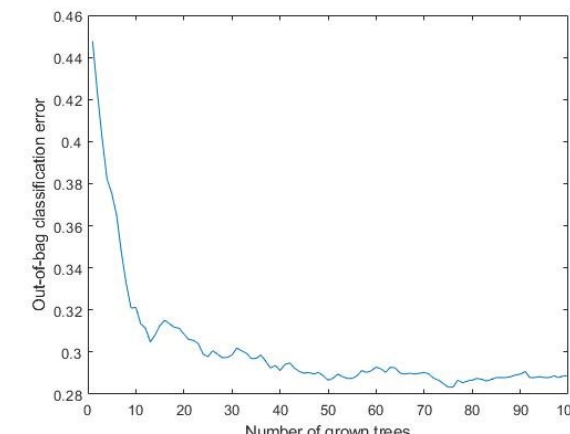


COMPARISON OF LOGISTIC REGRESSION AND RANDOM FORESTS ALGORITHMS FOR PREDICTION OF ASTHMA USING HEALTH SURVEY DATA																																																																		
Sergio Naval and Veronica Tuffrey City, University of London																																																																		
Problem statement and introduction																																																																		
<ul style="list-style-type: none">Our aim is to compare and critically evaluate the performance on a binary classification task of two popular machine learning (ML) algorithms, logistic regression (LR) and random forests (RF).Target variable is individuals' current asthma diagnosis, and predictors are a range of demographic, socio-economic, environmental and health-related variables at individual and household level.Our dataset is derived from the Health Survey for England (HSE) 2010 dataset [1]. HSE is an annual survey commissioned by the National Health Service to monitor trends in health. 2010 HSE includes 1,608 variables. This amount of variables present both an opportunity and a challenge: An opportunity to identify unknown complex relationships between asthma and variables, and a challenge to distinguishing relationships from noise.In England, asthma prevalence rate is among the highest in the world. Each year, asthma leads to an estimated 100 deaths, 12.7 million lost working days, and healthcare costs of £1 billion [2].A research priority for tackling the problem of asthma in the UK is creation of machine learning algorithms to predict children with the greatest risk of developing asthma [3, p.17].																																																																		
Dataset description and pre-processing																																																																		
<ul style="list-style-type: none">2010 HSE included a representative population sample of adults & children, and a “boost” sample of 2-15 y.o. children.Original dataset N = 14,112 records, 1,608 variables.Our binary target variable ASTHMA (currently diagnosed) included 1,364 of sampled individuals, prevalence = 9.7%. For pre-processing we: <ul style="list-style-type: none">deleted variables related to survey methodology;deleted variables directly related to target variable;recoded missing values (-ve numbers) to zero. Additionally, before applying logistic regression we: <ul style="list-style-type: none">standardised continuous and ordinal variables (n = 336);one-hot encoded string and nominal categorical variables;used Principal Components Analysis (PCA) to reduce dimensionality [4].	<ul style="list-style-type: none">Now the datasets had 1,150 variables for LR and 941 for RF. We then took a random sample of group without asthma to balance each dataset, so final N = 2,737 records.Graph below shows association of target variable (ASTHMA = 1 red) and deprivation status (most deprived 5). 	 <table><tr><th></th><th colspan="2">Asthma prevalence with..</th></tr><tr><th>Predictor</th><th>..predictor present</th><th>..predictor absent</th></tr><tr><td>Damp in house</td><td>10.5%</td><td>9.2%</td></tr><tr><td>Fungus in house</td><td>12.0%</td><td>9.1%</td></tr><tr><td>Passive smoker in house</td><td>13.5%</td><td>9.4%</td></tr><tr><td>Live in North-West of England</td><td>11.4%</td><td>9.4%</td></tr><tr><td>In richest income quintile</td><td>7.2%</td><td>10.3%</td></tr><tr><td>Live in village, hamlet or isolated dwelling</td><td>8.0%</td><td>9.9%</td></tr></table> <p>From preliminary analysis, we identify relationships between individual predictors and asthma:</p> <ul style="list-style-type: none">Graph above shows sample age distribution and asthma prevalence trend by age & genderTable above right shows association between asthma prevalence rate and examples of binary predictor variables.		Asthma prevalence with..		Predictor	..predictor present	..predictor absent	Damp in house	10.5%	9.2%	Fungus in house	12.0%	9.1%	Passive smoker in house	13.5%	9.4%	Live in North-West of England	11.4%	9.4%	In richest income quintile	7.2%	10.3%	Live in village, hamlet or isolated dwelling	8.0%	9.9%																																								
	Asthma prevalence with..																																																																	
Predictor	..predictor present	..predictor absent																																																																
Damp in house	10.5%	9.2%																																																																
Fungus in house	12.0%	9.1%																																																																
Passive smoker in house	13.5%	9.4%																																																																
Live in North-West of England	11.4%	9.4%																																																																
In richest income quintile	7.2%	10.3%																																																																
Live in village, hamlet or isolated dwelling	8.0%	9.9%																																																																
Summary of the two machine learning models, and their comparison																																																																		
Logistic Regression <ul style="list-style-type: none">LR assumes that log-odds of target class is a linear combination of features:$\ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \beta_1 X_i + \beta_0$LR provides probabilities for classifications. This enables adjustment of thresholds to best fit research question.To help avoid overfitting, regularization can be used - this is intended to reduce the model's variance without substantial increase in its bias.LR regularization is based on penalization of high weight (β) values: Ridge uses L2-norm ($L + \lambda \sum \beta_1^2 \dots$) and Lasso uses L1-norm ($L + \lambda \sum \beta_1 \dots$) where L is Negative Log Likelihood and λ is penalization parameter.	Random Forests <ul style="list-style-type: none">RF is an ensemble method based on decision trees (DT).Creates multiple independent DTs predictors from which predictions are aggregated. Independence of predictors is achieved using random sampling of Data – via bootstrap, and Features – via random choice.Predictions generated by averaging independent predictors reduces variance and thereby helps avoid overfitting.Individual DTs break feature space so subsets are “homogeneous” with respect to target variable. There are two common metrics of homogeneity:<ul style="list-style-type: none">“Gini impurity” $Gini(E)=1-\sum_{j=1}^c p_j^2$ (used by default in Matlab)“Entropy/Information gain” : $H(E)=-\sum_{j=1}^c p_j \log p_j$	<table><tr><th></th><th>Logistic Regression</th><th>Random Forests</th></tr><tr><td>Statistical assumptions? (e.g. linear data)</td><td>- Assumptions exist -> eg non-linear data needs transforming</td><td>+ No statistical assumptions – can handle multi-collinearity</td></tr><tr><td>Speed of training?</td><td>+ Faster than RF</td><td>- Slower than LR</td></tr><tr><td>Model explainability and interpretability of findings?</td><td>+ More easily explained and interpreted than RF</td><td>- Can be perceived as “black box” technique</td></tr><tr><td>Likelihood of overfitting?</td><td>- Overfitting & outliers a concern</td><td>+ Robust to overfitting and outliers</td></tr><tr><td>With large feature space..</td><td>- Performs less well than RF</td><td>+ Can handle many features</td></tr><tr><td>With many categorical variables..</td><td>- Performs less well than RF</td><td>+ Can handle many CVs</td></tr><tr><td>With low signal-to noise ratio and little data (i.e. a 'hard problem') ..</td><td>+ Performs better than RF</td><td>(RF has better performance on many medium-sized tasks)</td></tr></table>		Logistic Regression	Random Forests	Statistical assumptions? (e.g. linear data)	- Assumptions exist -> eg non-linear data needs transforming	+ No statistical assumptions – can handle multi-collinearity	Speed of training?	+ Faster than RF	- Slower than LR	Model explainability and interpretability of findings?	+ More easily explained and interpreted than RF	- Can be perceived as “black box” technique	Likelihood of overfitting?	- Overfitting & outliers a concern	+ Robust to overfitting and outliers	With large feature space..	- Performs less well than RF	+ Can handle many features	With many categorical variables..	- Performs less well than RF	+ Can handle many CVs	With low signal-to noise ratio and little data (i.e. a 'hard problem') ..	+ Performs better than RF	(RF has better performance on many medium-sized tasks)																																								
	Logistic Regression	Random Forests																																																																
Statistical assumptions? (e.g. linear data)	- Assumptions exist -> eg non-linear data needs transforming	+ No statistical assumptions – can handle multi-collinearity																																																																
Speed of training?	+ Faster than RF	- Slower than LR																																																																
Model explainability and interpretability of findings?	+ More easily explained and interpreted than RF	- Can be perceived as “black box” technique																																																																
Likelihood of overfitting?	- Overfitting & outliers a concern	+ Robust to overfitting and outliers																																																																
With large feature space..	- Performs less well than RF	+ Can handle many features																																																																
With many categorical variables..	- Performs less well than RF	+ Can handle many CVs																																																																
With low signal-to noise ratio and little data (i.e. a 'hard problem') ..	+ Performs better than RF	(RF has better performance on many medium-sized tasks)																																																																
Methods for model training and evaluation																																																																		
<ul style="list-style-type: none">We first extracted a dataset of 300 randomly sampled records to evaluate final out-of-sample model performance (Test set). The remaining 2,473 records were used for model training and hyper-parameter evaluation.Given the binary classification task and the balanced nature of the dataset we chose Accuracy as performance metric to evaluate our models. Accuracy is number of correct predictions divided by total number of predictions = $(TP + TN)/(TP + TN + FP + FN)$. We also used Receiver Operating Characteristic (ROC) curve where TP rate is plotted against FN rate. The model with larger “Area Under the Curve” (AUC) is the better one.We used Cross-Validation (CV) via grid search to evaluate combinations of hyper-parameters = regularization parameters for LR; and Number of Trees, Leaf Size and Number of Predictors Sampled for RF.There was high risk of over-fitting as our dataset has many features compared to number of records. We therefore used 10 fold CV instead of 5 or fewer folds to: 1) obtain a less variable evaluation of validation set performance and 2) give the models a greater number of samples during the training stage.Finally, we built LR and RF classification models using the optimal parameters identified using the training datasets, applied these models to the testing set, and compared the accuracies obtained.Analysis was done using MATLAB software (Release 2019b, MathWorks Inc., Natick, Massachusetts, US).																																																																		
Hypothesis statement																																																																		
<ul style="list-style-type: none">We expected both algorithms to perform well, with RF outperforming logistic regression because this is a complex dataset and LR imposes linear assumptions to the relationships between features and target variable.In LR the high number of features translates to a high number of parameters which would contribute to model variance and hence risk of overfitting. While penalization would be expected to improve out-of-sample performance in LR, we would still expect RF to more adequately deal with noise in the dataset. Additionally, published comparative studies indicate better performance of RF compared to LR, for example:<ul style="list-style-type: none">In a benchmarking experiment, RF performed better than LR in 69% of 243 datasets. Mean difference between RF and LR was 2.9% for accuracy, and 4.1% for AUC [5].In an evaluation of 179 classifiers on 121 datasets, RF classifiers were most likely to be the best. RF frequently out-performed even neural networks [6].																																																																		
Experimental findings on choice of parameters																																																																		
Logistic Regression <ul style="list-style-type: none">We used grid-search to evaluate the effect of penalization (λ) between 10^{-6} to 10^{-2}.For Lasso, best validation accuracy obtained with $\lambda=10^{-5}$ (left figure); for Ridge with $\lambda=10^{-4}$.We found that applying dimensionality reduction (PCA) to our dataset helps improve LR performance by 2%. In the PCA transformed dataset, Ridge regularization improved accuracy to 66.4% and Lasso to 66.1% (right figure shows medians obtained during grid search). 	Random Forests <ul style="list-style-type: none">We evaluated combinations of 3 hyper-parameters:<ul style="list-style-type: none">1) Number of Trees: Increases had little impact beyond about 25 trees. Out of bag classification error plateaued at around 50 Trees (see figure below)2) Leaf Size: This parameter had great impact. Best validation performance at around 30 (see figure right)3) Number of Predictors: This also had big impact. Performance continued to improve as number of predictors increased beyond 100, and only plateaued by 250 (see table right). 	 <table><tr><th></th><th colspan="6">Predictors Sampled -</th><th></th></tr><tr><th></th><th>5</th><th>15</th><th>50</th><th>100</th><th>250</th><th>500</th><th>Total</th></tr><tr><td>1</td><td>66.6%</td><td>68.3%</td><td>69.8%</td><td>70.6%</td><td>71.2%</td><td>71.0%</td><td>69.6%</td></tr><tr><td>5</td><td>66.7%</td><td>67.7%</td><td>69.7%</td><td>70.1%</td><td>71.2%</td><td>70.7%</td><td>69.3%</td></tr><tr><td>15</td><td>66.6%</td><td>68.0%</td><td>70.6%</td><td>70.9%</td><td>71.0%</td><td>71.4%</td><td>69.7%</td></tr><tr><td>30</td><td>65.6%</td><td>67.5%</td><td>69.6%</td><td>71.3%</td><td>71.7%</td><td>71.0%</td><td>69.5%</td></tr><tr><td>60</td><td>65.7%</td><td>67.3%</td><td>69.3%</td><td>70.1%</td><td>70.1%</td><td>70.7%</td><td>68.9%</td></tr><tr><td>Total</td><td>66.2%</td><td>67.8%</td><td>69.8%</td><td>70.6%</td><td>71.0%</td><td>71.0%</td><td>69.4%</td></tr></table>		Predictors Sampled -								5	15	50	100	250	500	Total	1	66.6%	68.3%	69.8%	70.6%	71.2%	71.0%	69.6%	5	66.7%	67.7%	69.7%	70.1%	71.2%	70.7%	69.3%	15	66.6%	68.0%	70.6%	70.9%	71.0%	71.4%	69.7%	30	65.6%	67.5%	69.6%	71.3%	71.7%	71.0%	69.5%	60	65.7%	67.3%	69.3%	70.1%	70.1%	70.7%	68.9%	Total	66.2%	67.8%	69.8%	70.6%	71.0%	71.0%	69.4%
	Predictors Sampled -																																																																	
	5	15	50	100	250	500	Total																																																											
1	66.6%	68.3%	69.8%	70.6%	71.2%	71.0%	69.6%																																																											
5	66.7%	67.7%	69.7%	70.1%	71.2%	70.7%	69.3%																																																											
15	66.6%	68.0%	70.6%	70.9%	71.0%	71.4%	69.7%																																																											
30	65.6%	67.5%	69.6%	71.3%	71.7%	71.0%	69.5%																																																											
60	65.7%	67.3%	69.3%	70.1%	70.1%	70.7%	68.9%																																																											
Total	66.2%	67.8%	69.8%	70.6%	71.0%	71.0%	69.4%																																																											
Experimental findings from optimised models, and critical evaluation																																																																		
Logistic Regression <ul style="list-style-type: none">Our comparison (right) shows LR underperformed by 5.7% for accuracy and 5.1% for AUC compared with RF, consistent with our initial hypothesis. Underperformance is larger than the average in the study by Couronné et al [5], as expected given complexity of dataset (e.g. interactions between variables).The underperformance of LR could be related to the dataset not meeting the model's linear separability assumption.The difference would likely have been larger if dimensionality reduction via PCA and regularization had not been applied.The findings point to difficulties of LR to adequately fit a highly noisy dataset with high ratio of features to records.There was only a small difference between validation and training accuracy (2.7%) indicating model was not over-fitting.	 <table><tr><th>LR*</th><th>Results</th><th>RF**</th></tr><tr><td>69.1%</td><td>Training</td><td>78.7%</td></tr><tr><td>66.4%</td><td>Validation</td><td>71.6%</td></tr><tr><td>65.3%</td><td>Test</td><td>71.0%</td></tr><tr><td>73.3%</td><td>Test AUC</td><td>78.4%</td></tr></table> <p>* Using Ridge regression with $\lambda=0.01$ ** Using 100 trees, 250 predictor variables, and minimum leaf size of 30</p>	LR*	Results	RF**	69.1%	Training	78.7%	66.4%	Validation	71.6%	65.3%	Test	71.0%	73.3%	Test AUC	78.4%	Random Forests <ul style="list-style-type: none">RF had good performance compared to LR, perhaps because RF can cope better with complex relationships between variables (e.g. interactions).The ROC curve shows RF achieves higher sensitivity than LR for same specificity (1 – false +ve rate) when specificity is >50%, indicating RF identifies a higher proportion of cases, important for health applications.We found performance varied greatly according to hyper-parameters, and optimal choices related to our dataset characteristics. For example, while default number of predictors would be c.40 (= square root of number of features [7]), we found validation accuracy plateaued at c. 250 features, consistent with our dataset having many non-relevant feature variables.While both RF and LR models are quite small and quick, the higher performance of RF is paired with a larger model (2051 KB vs. 4 KB) and greater training times (mean 6 seconds vs. 0.1 seconds).																																																	
LR*	Results	RF**																																																																
69.1%	Training	78.7%																																																																
66.4%	Validation	71.6%																																																																
65.3%	Test	71.0%																																																																
73.3%	Test AUC	78.4%																																																																
Lessons learnt and suggestions for future work																																																																		
<ul style="list-style-type: none">For our dataset, accuracy from application of RF was 5.7% higher than with LR, and sensitivity values for RF were higher or equal to those for LR for the same specificity. These results are consistent with non-linear or complex associations existing between some of the variables in HSE and asthma. RF has been able to distinguish these associations from the noise present on the dataset which shows that the method is well suited for complex and medium datasets where LR assumptions might not be met.<u>Suggestions for future work:</u><ul style="list-style-type: none">Test whether investing significant more time at the data pre-processing stage to delete non-relevant variables would reap significant benefits with respect to the predictive power of LR and RF;Compare findings to those obtained using Gradient Boosting and Support Vector Machines (SVM). We would expect both approaches to perform well for a binary classification task on a large noisy dataset [5, 6];Use our findings as basis of project relating to explainability in ML for public health practitioners, that is, how best to explain RF findings and describe associations between variables and asthma.In summary, since RF was able to predict presences of asthma well using a fairly unclear dataset, our study implies machine learning can be helpful to predict complex health-related associations, and thus help inform policy-making with respect to asthma and other debilitating conditions in the UK.																																																																		

[1] UK Data Service. 2019. *Health Survey for England, 2010* [https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=6986].

[2] Joint Health Surveys Unit. *Health Survey for England 2010 Respiratory health, Summary of key findings*. London: National Centre for Social Research and UCL; 2011. [https://files.digital.nhs.uk/publicationimport/pub03xxx/pub03023/heal-surv-eng-2010-resp-heal-summ-rep.pdf].

[3] Abel J, Poinasamy K, Takhar P. *Asthma still kills: Urgent priorities for the international research community to treat, prevent and cure asthma*. London: Asthma UK; 2019. [https://www.asthma.org.uk/60a27fe6/globalassets/campaigns/publications/ae-report-final-approved.pdf]

[4] Aguilera AM, Escabias M, Valderrama MJ. Using principal components for estimating logistic regression with high-dimensional multicollinear data. *Computational Statistics & Data Analysis* 2006; 50:1905-1924. [https://doi.org/10.1016/j.csda.2005.03.011]

[5] Couronné R, Probst P, Boulesteix A-L. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics* 2018; 19:270. [https://doi.org/10.1186/s12859-018-2264-5]

[6] Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research* 2014; 15:3133-81. [http://jmlr.org/papers/volume15/delgado14a/delgado14a.pdf]

[7] Scornet E. Tuning parameters in random forests. *ESAIM: Proceedings and Surveys* 2017; 60:144-162. [https://doi.org/10.1051/proc/201760144]