Follow the Trend

Krist Papadopoulos

CSC2515 Introduction to Machine Learning
Fall 2017 Project Paper
University of Toronto

Abstract. In this paper, lattice regression models with monotonicity biases were compared to lasso and random forest regression on the baseline Boston housing dataset and an outlier filtered version of the same dataset. It was found that lattice regression models such as calibrated lattice and random tiny lattice produced lower average mean squared error predictions and less variable predictions from 5-fold cross validation than both lasso and random forest regression. The specification of a prior monotonicity bias in the lattice regression models based on linear correlations of features with the target variable did not lead to improved predictions over the baseline lattice regression models but it did demonstrate the lowest average model prediction standard deviation on the baseline dataset.

1 Introduction

The ability to extract interpretable predictions from the data distributions is important for application of machine learning algorithms in fields such as law, medicine and finance. Regression techniques for predicting real valued outputs from a set of known targets and input features, vary in prediction accuracy, expressivity and interpretability. More biased and interpretable techniques such as linear regression make strong assumptions about the relationship of output target to the input features. Linear regression may indicate a trend in the output but its prediction accuracy to unseen data can be limited due to data variability and deviations from the linear model assumption. Less biased techniques with less interpretability such as deep neural networks (DNN) avoid the assumptions required for linear regression and can accurately approximate functions between target and features although this expressivity does not necessarily generalize well to unseen data especially if the noise in the training dataset is learned (Zhang et al., 2017) which poses potential problems for learning from tabular datasets where the difference between the trend or signal and noise in the data may not be clear.

The purpose of this paper is to investigate and compare the performance of the following interpretable regression that have different levels of biases and expressivity in fitting the mapping between features and target : lasso regression (Tibshirani, 1996), random forest regression (Breiman, 2001), lattice regression (Garcia and Gupta, 2009 and Garcia et al., 2012), monotonic calibrated lattice regression (Gupta et al., 2016) and random tiny lattice (RTL) (Canini et al., 2016) on a tabular dataset that would be encountered in real world regression applications. Lattice regression implementations have unique features such as calibrated inputs, linear function interpolation and monotonicity bias that could be useful in making predictions where some degree of linearity and function smoothness (i.e. following a trend as opposed to variable non-linear changes) is expected in the target.

The hypothesis in this paper is that the target (i.e. house prices) in the housing dataset (Boston Housing – Harrison and Rubinfeld, 1978) has monotonically or partial monotonically increasing/decreasing correlations with features which form a trend in the target variable in amongst the variability introduced from varying features, observations and nonlinearities. If a trend of the target to a feature can be established, either by correlation analysis or from experience then it is expected that specification of this relationship (i.e. positive or negative correlation of feature(s) with target) a prior in the model will bias the model to be able to generalize better to unseen data from the data distribution as opposed to models that are too rigid in their hypothesis (i.e. Linear Regression) or may get influenced more by the variability in the data (i.e. Random Forest Regression). To test this hypothesis, the dataset was evaluated using lattice regression with monotonicity bias and compared to linear regression with L1 regularization (i.e. lasso regression) and random forest regression. See Figure 1 below for a schematic of the experiments performed in this paper.

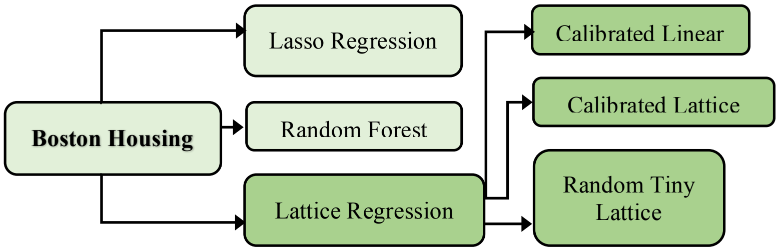

Figure 1. Outline of experiments performed in this paper

It is expected that specifying monotonicity biases along with the other features of lattice regression such as piece-wise linear calibration of features and lattice function interpolation will reduce overfitting and produce better generalization performance (i.e. lower mean squared error on the test data sets) compared to lasso and random forest regression on the same dataset. The housing dataset was selected as an example from finance where the interpretability of the results and model may be important for designing processes or applications (i.e. home value for proper-ty tax assessment). Other types of datasets in finance such as credit card and loan default risk were not tested. Another limitation of this paper is that DNN and recent developments in deep lattice networks (DLN) (You et al., 2017) were not tested. The limitations of this work are discussed further in Section 3.

1.1 Related Work

The experiments in this paper continue from the lattice regression experiments performed in prior work. A lattice is a set of *m* fixed nodes that can be represented in a *d* dimensional grid. The lattice is used with the training data to estimate a function class that linearly interpolates lattice outputs which minimizes empirical risk. Garcia and Gupta (2009) developed the empirical risk formulation of the lattice regression algorithm with graph Laplacian regularization to enforce smoothness of the lattice outputs by penalizing the squared difference in lattice output values for adjacent nodes and with a global bias to encourage linear extrapolation of the learned function beyond the boundary of the lattice grid cells. They tested the performance of this approach on randomly generated functions, real geo-spatial data and real-data colour management tasks.

Garcia et al. (2012) continued investigation of lattice regression for colour management tasks and omnidirectional super-resolution for visual homing. In this work a new lattice regularization called the graph Hessian was introduced. The graph Laplacian regularization that was introduced by Garcia and Gupta (2009) was argued to be sub-optimal for many applications since the enforcement of lattice node smoothness would potentially penalize linear trends in the data. The graph Hessian penalizes the second order difference in each dimension of the lattice, summed over the *d* dimensions.

Gupta et al. (2016) introduced monotonic calibrated interpolated look-up tables for lattice regression. The motivation for this work was to find a method to guarantee the monotonicity of the learned function for some inputs for applications where a blackbox model may not be acceptable but while not trading off for accuracy. They argued that a key interpretability issue for machine learning models was whether the learned model can be guaranteed to be monotonic with respect to some features. A new form of regularization called torsion regularization was also introduced to penalize twists in the lattice sides to promote the linearization of the learned function. In this work, they demonstrated large scale learning of monotonic lattice models with up to 12 features on datasets such as business matching, ad-query matching, rendering classifiers and video ranking.

Canini et al. (2016) further explored the applications of monotonic lattice regression and demonstrated that monotonic ensembles of lattices can be learned. The random tiny lattice (RTL) was introduced for a random subset of features selected for each lattice uniformly and independently without replacement from the set of total features. The advantage of training an ensemble of lattices was to reduce evaluation time and memory usage especially with larger features sized datasets. Experiments showed that the test accuracy of monotonic ensembles of lattice outperformed random forest on a regression task of scoring the quality of a candidate for a matching problem.

The details of the experiments performed in this paper are described below.

2 Experiments

The regression models were evaluated on the Boston housing (Harrison and Rubinfeld, 1978) dataset which is summarized in Table 1. The dataset was divided with 5 fold cross-validation for training and prediction with different hyperparameters. The lasso and random forest regression models were implementations from Scikit-learn (Pedrogosa et al., 2011) and the lattice regression models were implementations from TensorFlow (Abadi et al., 2015) and TensorFlow Lattice (Gupta et al., 2017).

Dataset	Features	Target	Observations
Boston	13	House Price	506

Table 1. Summary of the dataset

The target is approximately log normally distributed, although there were 16 observations at the maximum house price of 50 that did not follow the distribution. These observations may be outliers in the dataset but there were retained for analysis. The regression algorithms were tested on the total dataset with the observations at 50 and they were also tested with the potential outliers at 50 filtered out to determine the impact on prediction performance. The average mean squared error (MSE) and the sample standard deviation (SD) were reported from the 5-fold cross validation. The linear correlations between features and the target variable are depicted in Figure 2. From Figure 2, the feature RM (i.e. number of rooms) was selected to test positive monotonicity with the target variable in the lattice model, since it has the highest positive correlation (0.7). It was expected that houses with more rooms would consistently have a higher price. The feature LSTAT (i.e. proportion of the population with education) was selected to test negative monotonicity with the target variable in the lattice model since it has the highest negative correlation (-0.74). It was expected that there would be a consistent inverse relationship with house prices and LSTAT. The features of the dataset were all continuous, real-valued which were standardized for the lasso and random forest regression by subtracting the mean and dividing by the standard deviation. The features were not standardized for the lattice models because the lattice models normalize the features through a process called calibration. No other feature pre-processing was performed.

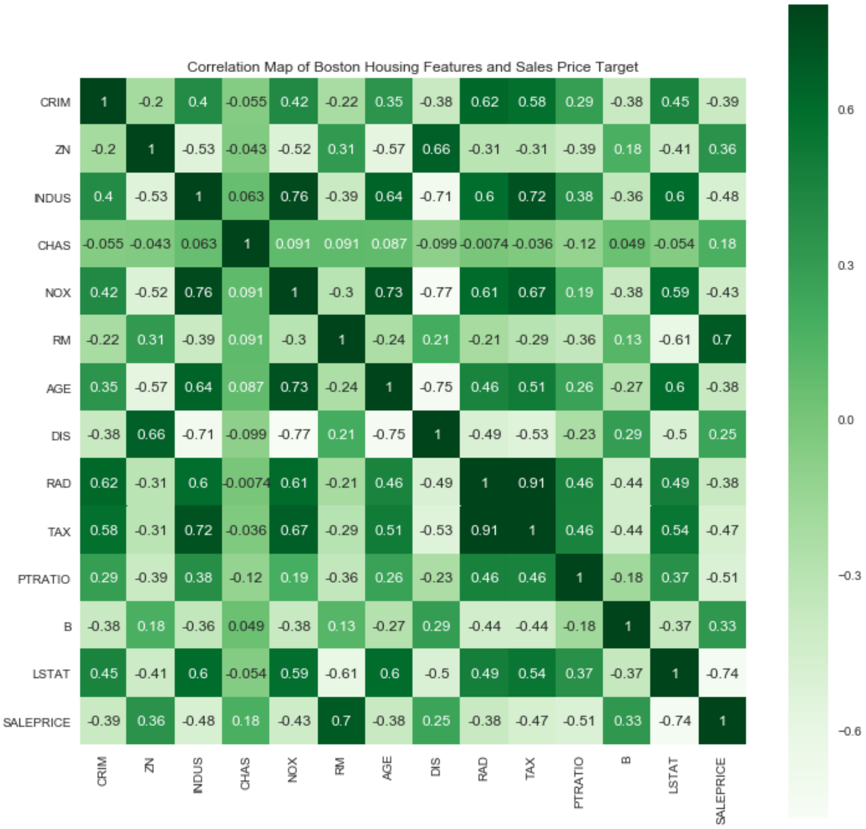

Figure 2. Linear correlation map of Boston housing features and sales price

2.1 Lasso Regression

Lasso regression (i.e. linear regression with L1 regularization) was applied to the total dataset and the filtered dataset. The results are presented in Table 2. In Figure 3, the 5 prediction distributions (i.e. curves) from the 5-fold cross validation are plotted over the target distribution which is solid green. Figure 3 and Figure 4, show that the target predictions significantly deviated from the target distribution. Since the target it not a linear function of the features, the lasso regression model bias was to too rigid and therefore it had limited capability to reflect the changes in the target distribution. The removal of the observations at 50, resulted in more accurate predictions, although the predictions were still highly variable in the tail of the distribution as shown in Figure 4.

Lasso Regression	Average MSE	SD
All Data, L1 = 0.02	23.48	0.84
Filtered, L1 = 0.01	15.16	1.44

Table 2. Summary of the lasso regression results

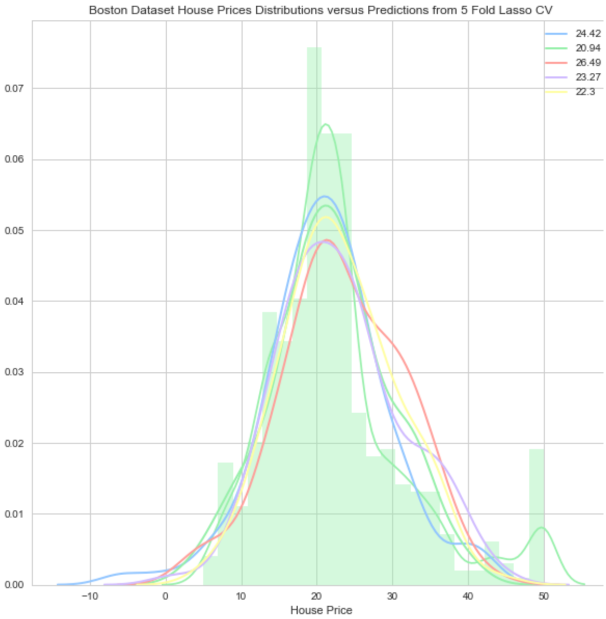

Figure 3. Lasso Regression (All Data)

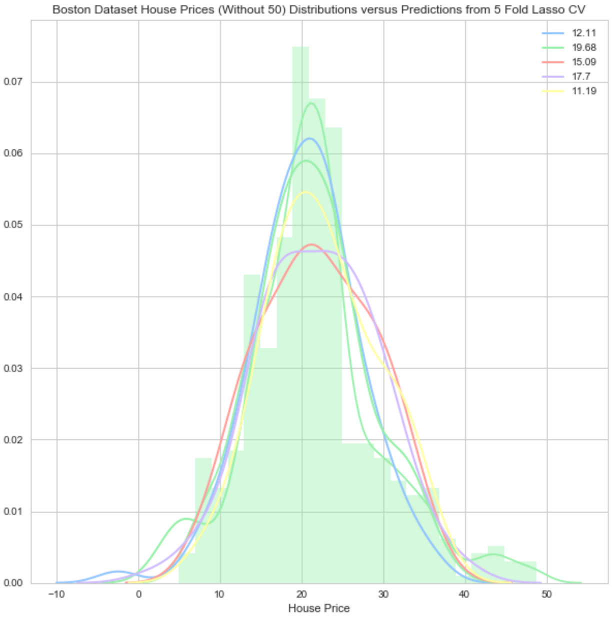

Figure 4. Lasso Regression (Filtered)

2.2 Random Forest Regression

Random forest regression was applied to the total dataset and the filtered dataset. The results are presented in Table 3. The hyperparameters of the random forest model (i.e. number of trees, and number of features to randomly select) were tuned using grid search cross validation. Random forest exhibited better performance than lasso regression on the total dataset and on the filtered dataset. Random forest does not make the linear assumption that lasso regression does which allows it to fit a more variable distribution. The fit of the random forest model was more centered and stable as it did not shift significantly for each trial as shown in Figure 5. The tail of the distribution which was not accurately predicted by the lasso regression was better fit by random forest, although the presence of the outliers at 50, still impacted the predictions as depicted in the differences between Figure 5 and Figure 6. The removal of the outliers shown in Figure 6, resulted in improved predictions but significant variability still existed in the predictions from the tail of the distribution.

Random Forest Regression	Average MSE	SD
All Data: Trees = 500, Features = 7	10.10	0.76
Filtered: Trees = 400, Features = 6	7.31	0.68

Table 3. Summary of the random forest regression results

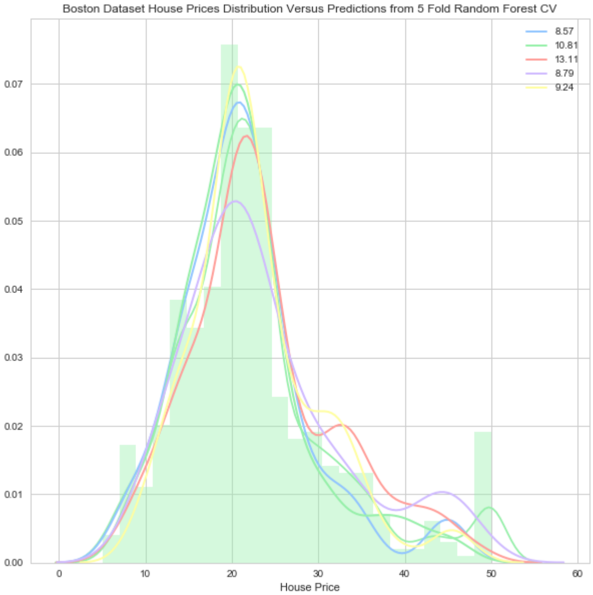

Figure 5. Random Forest Regression (All Data)

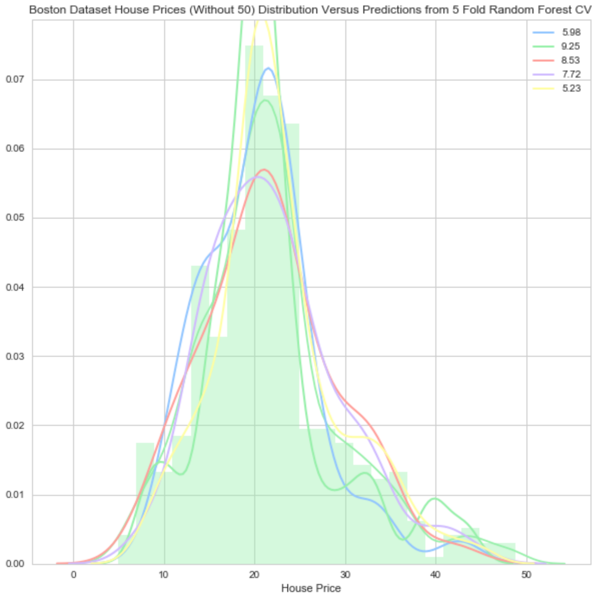

Figure 6. Random Forest Regression (Filtered)

2.3 Lattice Regression

The calibrated linear model (Gupta et al., 2017), learns a 1-D mapping for each feature using 1-D lattices and then combines the feature transformations linearly. The 1-D lattice has a range from 0 to 1 to which the feature values are calibrated or normalized using piece-wise linear functions with lattice keypoints. The number of keypoints used in the calibration of the features is a hyperparameter with more keypoints offering more flexibility in representing the feature range into the lattice range. In this paper 10 keypoints were used for all lattice models. The results for the calibrated linear model are presented in Table 4. The calibrated linear model performed better than lasso regression but not as well as random forest. In the experiments, the usage of calibration regularization (L1, L2 = [0.001, 0.01, 0.05]) and monotonicity biases in two features (RM and LSTAT) that exhibited the strongest linear correlation with the target did not significantly change the performance of the models. From Figure 7 and Figure 8, the calibrated linear model exhibited less variable predictions as compared to lasso regression and was not as significantly impacted by the outlier at 50. The calibrated linear model cannot model non-linear feature interactions and therefore it was not as expressive as the random forest model in fitting the distribution although the calibration of the features reduced variance in the predictions and did not over bias the predictions like lasso regression did.

Calibrated Linear Model (1-D Lattices)	All Data		Filtered	
	Average MSE	SD	Average MSE	SD
Baseline (no monotonicity or regularization)	15.11	0.97	13.00	0.49
Baseline with L1 regularization	15.08	1.18	12.89	0.48
Baseline with L2 regularization	14.77	1.13	11.96	0.47
Monotonicity in RM	14.65	1.22	13.14	0.49
Monotonicity in RM and LSTAT	14.55	1.21	13.08	0.48
Monotonicity in RM and LSTAT with L1 reg.	14.53	1.20	13.06	0.46
Monotonicity in RM and LSTAT with L2 reg.	14.29	1.14	11.99	0.97

Table 4. Summary of the calibrated linear results

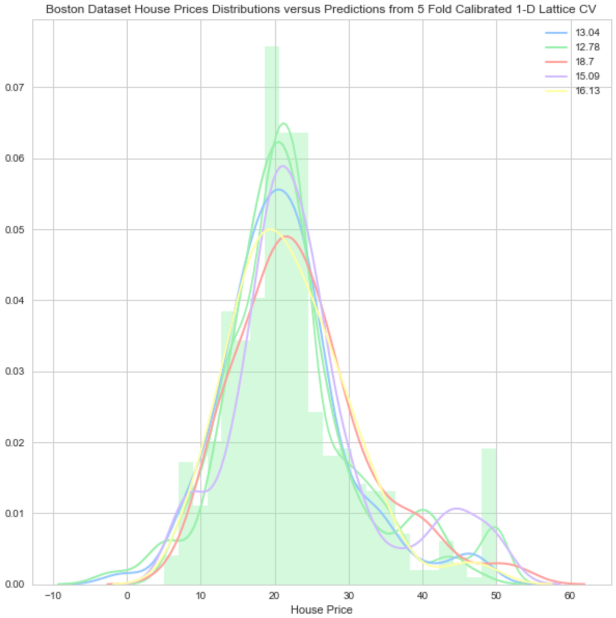

Figure 7. Calibrated Linear Baseline (All Data)

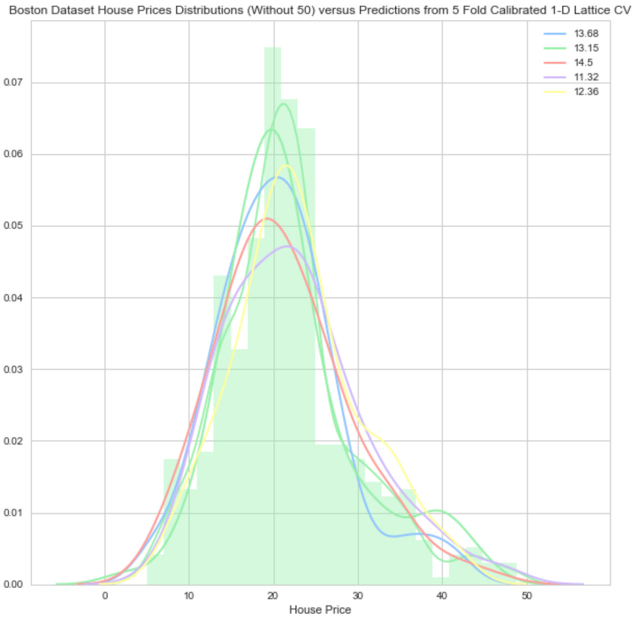

Figure 8. Calibrated Linear Baseline (Filtered)

An extension to the calibrated linear model is the calibrated lattice (Gupta, 2017), where all the features are calibrated and then combined into a layered lattice that is learned. For a calibrated lattice, the number of parameters (i.e. lattice nodes) to be learned is exponential in the number of features. Therefore, training calibrated lattices greater than 10 features are not recommended (Gupta, 2017). The advantage of the calibrated lattice over the calibrated linear model (1-D lattice) is that non-linear interactions between all the features can be learned in the layered lattice. The calibrated lattices used in the experiments were trained on all 13 features. To train models with larger feature sets, it was shown (Canini et al, 2016) that ensembles of lattices can be trained and linearly combined. In this paper, the random tiny lattice model (Gupta, 2017) was trained with 100 two-layer lattices with each having 4 features randomly selected from the total 13 features. Both the calibrated lattice and the random tiny lattice were trained with a mini-batch size of 100 for 15000 training steps. Similar to the calibrated linear model, L1/L2 calibration regularization was used for both models to test the effect of regularizing the fit of each feature range to the lattice range since only 10 keypoints were used in the feature transformations. From Table 5, the calibrated lattice model had the lowest MSE on the total dataset with L1 calibration regularization. The RTL model achieved lowest MSE on the filtered dataset with L1 calibration regularization. It was expected that the calibrated lattice would perform better on the total dataset with the outliers since it learned non-linear interactions between all the features where in the RTL, the lattices in the ensemble do not incorporate all features per lattice. As demonstrated in Figure 7 and Figure 8, each prediction from the calibrated lattice model was followed the target distribution accurately with low variance, and was not significantly influenced by the outliers or tail of the distribution. Overall, the performance of both the calibrated lattice and the RTL with calibration regularization were comparable across both datasets in terms of MSE and SD. Both models significantly outperformed the random forest regression predictions.

The hypothesis that monotonicity biases in features RM and LSTAT would improve the predictions was not realized with either the calibrated lattice or RTL models. With calibration regularization, the predictions with monotonicity biases improved from the baseline monotonicity predictions as shown in Table 5, but these predictions were comparable to those from the calibrated lattice and RTL without the monotonicity biases. It was observed that including LSTAT negative monotonicity with RM positive monotonicity bias in the RTL model slightly reduced the prediction error which suggested that the inclusion of only RM positive monotonicity bias was maybe too strong an assumption for the dataset. The training of the calibrated lattice model with both RM and LSTAT monotonicity biases was not completed because of the runtime encountered and therefore further experiments with the calibrated lattice and monotonicity biases were not pursued. It was also observed on the total dataset that the standard deviation of the RTL model predictions with RM/LSTAT monotonicity and L1 calibration regularization was the lowest achieved for all experiments. When the outliers were removed a similar standard deviation was not achieved with the same model. This finding suggested that the monotonicity of both RM and LSTAT was effective in regularizing the model predictions in the presence of the outlier observations because these observations were consistent in that they had greater number of house rooms with high house prices which was consistent with the RM positive monotonicity bias and the LSTAT negative monotonicity bias. It is not understood why this result was only achieved with L1 calibration regularization and not with L2 calibration regularization.

Calibrated Lattice/Random Tiny Lattice (RTL)	**All Data**		**Filtered**	
	Average MSE	**SD**	**Average MSE**	**SD**
Calibrated Lattice baseline (no monotonicity or regularization)	4.15	0.28	4.19	0.56
Calibrated Lattice with L1 regularization	3.87	0.52	3.49	0.27
Calibrated Lattice with L2 regularization	4.09	0.53	5.13	0.32
Calibrated Lattice with monotonicity in RM	6.86	1.88	5.97	0.70
RTL baseline (no monotonicity or regularization)	5.19	0.36	4.05	0.39
RTL baseline with L1 regularization	5.30	0.49	2.97	0.24
RTL baseline with L2 regularization	4.23	0.22	3.71	0.21
RTL with monotonicity in RM	6.24	0.47	4.38	0.37
RTL with monotonicity in RM with L1 regularization	4.87	0.36	3.76	0.29
RTL with monotonicity in RM with L2 regularization	5.47	0.44	4.10	0.26
RTL with monotonicity in RM and LSTAT	5.47	0.31	3.97	0.21
RTL with monotonicity in RM and LSTAT with L1 reg.	4.32	0.09	3.67	0.28
RTL with monotonicity in RM and LSTAT with L2 reg.	4.69	0.36	3.67	0.35

Table 5. Summary of the calibrated lattice and random tiny lattice results

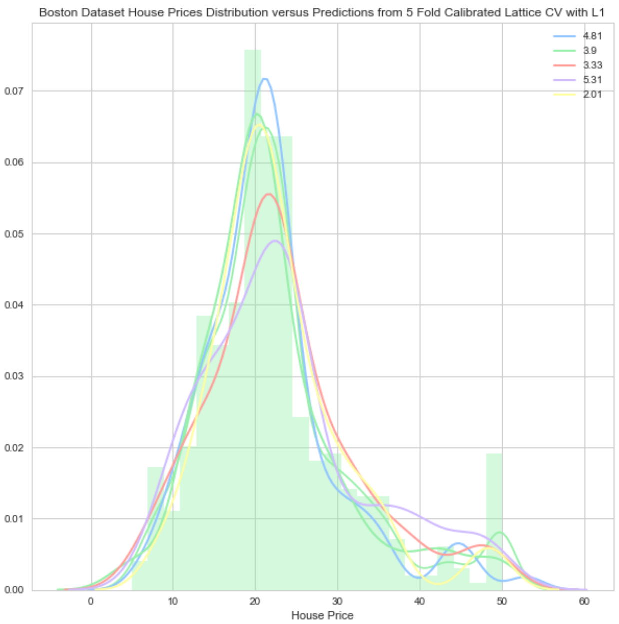

Figure 7. Calibrated Lattice with L1 reg. (All Data)

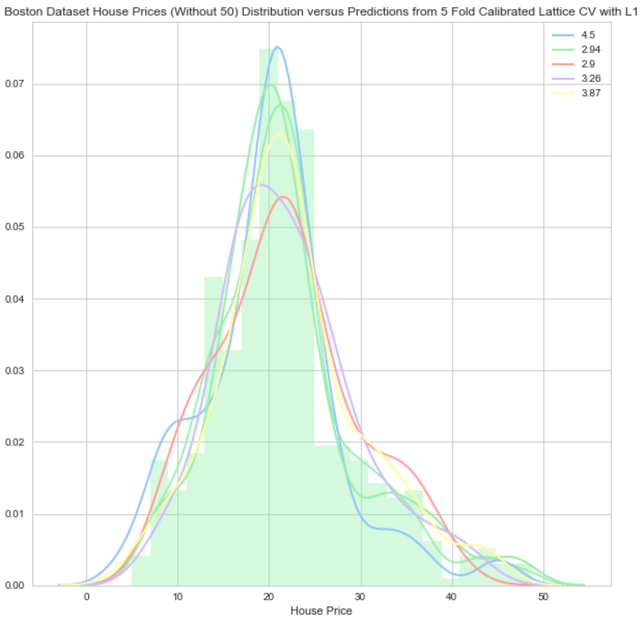

Figure 8. Calibrated Lattice with L1 reg. (Filtered)

3 Limitations

This paper only explored the performance of the lattice regression models with monotonicity on one dataset. Further analysis on different types of tabular datasets is required to examine the capability of lattice regression models to fit different data distributions and make accurate predictions on unseen data. The features selected for monotonicity bias were only identified through linear correlation of the baseline features and the target. A different selection technique either through feature engineering, or examining only certain ranges of monotonicity with regularization were not examined in this paper. The comparisons made in this paper were limited only to lasso and random forest regression. Lattice regression can be compared to models such as DNN and DLN, to further examine the trade-offs between bias, variance and model interpretability. The lattice regression implementation used in this paper (Gupta et al., 2017) incorporated many hyperparameters (e.g. keypoints, number of lattice layers, number of lattices in RTL ensemble, number of features in RTL) that were not varied in the experiments and different forms of regularization in the calibrated lattice and RTL models (i.e. global bias and Laplacian regularization (Garcia and Gupta, 2009), torsion regularization (Gupta et al., 2016)) were not examined. Further work is required to understand how these hyperparameters and regularizations impact modelling with different data distributions.

4 Conclusions

In this paper, it was shown that calibrated lattice and RTL models were effective in significantly improving the prediction performance (i.e. lower average MSE and lower sample standard deviation over 5-fold cross validation) compared to lasso and random forest regression on the baseline and filtered Boston housing dataset. The calibrated linear (1-D lattice) model was not as effective as the calibrated lattice and random tiny lattice because it cannot model non-linear feature interactions in the dataset, although the calibration of features, using piece-wise linear functions resulted in better predictions than lasso regression. The hypothesis that prediction accuracy would be improved by incorporating monotonicity biases into the lattice models from linear trends observed in the dataset was not observed for the features tested in this paper (RM and LSTAT). There was a finding where the monotonicity biases may have been effective in reducing the standard deviation RTL model predictions on the total dataset since the outliers in the dataset followed the RM and LSTAT monotonicity biases that were incorporated. In other datasets where the target response is expected to be more linear over a range or there is more noise in the data, monotonicity bias might prove to be more effective in regularizing the predictions from overfitting to noise in the data and thus produce improved predictions. The lattice regression models used in this paper proved to be effective without monotonicity bias, adding biases through piece-wise linear calibration of features, lattice function interpolation and calibration regularization. These linearizing lattice model features combined with the ability to model non-linear feature interactions in the lattice produced more accurate and less variable results compared to lasso and random forest regression with and without the presence of outliers in the data.

References

Abadi et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, *arXiv preprint*, 2015. URL https://arxiv.org/pdf/1603.04467

Leo Breiman. Random Forests. *Machine Learning*, vol. 45, no. 1, pp 5-32, 2001.

Kevin Canini, Andrew Cotter, Maya Gupta, Mahdi Milani Fard, Jan Pfeifer. Fast and Flexible Monotonic Functions with Ensembles of Lattices. In *Advances in Neural Information Processing Systems*, 2016.

Eric Garcia, Raman Arora, Maya R. Gupta. Optimized Regression for Efficient Function Evaluation. *IEEE Transactions on Image Processing,* 2012.

Eric Garcia, Maya Gupta. Lattice Regression. In *Advances in Neural Information Processing Systems*, 2009.

D. Harrison, D. L. Rubinfeld. Hedonic Housing Prices and the Demand for Clean Air. *Journal of Environmental Economics and Management*, 5, pp. 81-102., 1978.

Maya Gupta, Jan Pfeifer, Seungil You. Tensorflow Lattice: Flexibility Empowered by Prior Knowledge. *Google Research Blog*, 2017. URL https://research.googleblog.com/2017/10/tensorflow-lattice-flexibility.html

Maya Gupta, Andrew Cotter, Jan Pfeifer, Konstantin Voevodski, Kevin Canini, Alexander Mangylov, Wojciech Moczydlowski, Alexander van Esbroeck. Monotonic Calibrated Interpolated Look-Up Tables. *Journal of Machine Learning Research*, 2016.

Pedregosa et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 12, pp. 2825-2830, 2011.

Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society.* Series B (Methodological), vol. 58, no. 1, pp. 267–288, 1996.

Seungil You, David Ding, Kevin Canini, Jan Pfeifer, Maya R. Gupta. Deep Lattice Networks and Partial Monotonic Functions. In *Advances in Neural Information Processing Systems*, 2017.

Chiyaun Zhang, Sammy Bengio, Moritz Hardt, Benjamin Recht, Oriol Vinyals. Understanding Deep Learning Requires Rethinking Generalization. In *International Conference on Learning Representations*, 2017a.

Appendix

The code and results from this paper are available in the following repository:

https://github.com/kristpapadopoulos/CSC2515_Fall_2017_Paper