Bayesian model validation metrics in retail datasets

Leevi Rönty

School of Science

Bachelor's thesis
Espoo 14.10.2020

Supervisor

Prof. Fabricio Oliveira

Advisors

DSc (Tech.) Mikko Ervasti

DSc (Tech.) Paavo Niskala

Aalto University
School of Science

Aalto University
School of Science

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi
Abstract of the bachelor's thesis

Author Leevi Rönty

Title Bayesian model validation metrics in retail datasets

Degree programme Engineering Physics and Mathematics

Major Mathematics and Systems Sciences **Code of major** SCI3029

Teacher in charge Prof. Fabricio Oliveira

Advisors DSc (Tech.) Mikko Ervasti, DSc (Tech.) Paavo Niskala

Date 14.10.2020 **Number of pages** 0 **Language** English

Abstract

Multiple methods for model selection exist, but when applied to Bayesian models in a business setting, none of them is clearly better than the others. Not all metrics take into account the Bayesian nature of the models. Some methods require the model to be fitted multiple times with different data, but this might cause issues if the models are slow to fit. availability of data points may also be limited, hindering the applicability of methods requiring hold-out sets.

In this thesis we will study five popular model selection methods. The chosen metrics are Akaike information criterion (AIC), Deviance information criterion (DIC), Watanabe-Akaike information criterion (WAIC), mean absolute percentage error (MAPE) and 10-fold cross validation. We form five different models

Aalto University
School of Science

Aalto-yliopisto, PL 11000, 00076 AALTO
www.aalto.fi
Tekniikan kandidaatintyön tiivistelmä

Tekijä Leevi Rönty

Työn nimi Bayeslaisten mallien validointi vähittäismyynnin aineistoissa

Koulutusohjelma Teknillinen fysiikka ja matematiikka

Pääaine Matematiikka ja systeemitieteet **Pääaineen koodi** SCI3029

Vastuuopettaja Prof. Fabricio Oliveira

Työn ohjaajat DSc (Tech.) Mikko Ervasti, DSc (Tech.) Paavo Niskala

Päivämäärä 14.10.2020 **Sivumäärä** 0 **Kieli** Englanti

Tiivistelmä

Avainsanat Bayeslaiset mallit, mallin validointi, informaatiokriteeri, Facebook Prophet

Contents

1 Introduction

Businesses need forecasting to succeed. Correctly forecasting demand for goods is essential for any retail business, as shortages can cause loss of sales. As companies collect increasingly more data, the possibilities of forecastable subjects and possible features increase dramatically. However, one cannot just include more features in the model and expect it to perform well. Furthermore, not all models are suitable for all forecasting tasks. The number of possible models is endless, but most of them will perform poorly. To find the useful ones, we must be able to measure the goodness of a model. Being able to objectively compare models allows for systematic approaches of model selection to be adopted.

After **?**, model validation is an integral part of a robust modelling framework. Bayesian models are a relatively novel model type, which has gained considerable popularity in recent years, as computational power of modern computers has kept rising. As a relatively new method, not too many validation metrics have been proposed with these kinds of models in mind.

The current state of model validation for Bayesian models can be described as "unsatisfactory". Information criteria that try to predict out of sample fitness can have strong biases, and some may not take into consideration the nature of Bayesian models and distributions of parameters. Cross validation can be extremely computationally intensive. Additionally, most information criteria are based on a loss function proportional to the root mean square of errors, but that may not always be the most useful function to optimise for. In some businesses, the consequences of errors in forecasts can be described better with the mean absolute percentage error. All in all, there does not exist a clearly better method for solving all Bayesian model validation problems.

This thesis attempts to study the usefulness of popular model selection metrics when applied on retail datasets. How can the metrics be applied when the number of past observations is very limited? Does it significantly affect the performance or applicability of some metrics? Retail data can also be very detailed. Low-level modelling can lead to a very high model count or very complex models. How rapid are the evaluation methods? This can significantly impact the usefulness of the metric. As the models are applied to Bayesian models, there is always some randomness involved as the sampling process is not deterministic. Will this play a significant role in the stability of the metrics or will the results be always consistent? By answering these questions, we try to conclude about which metrics are useful in which circumstances in business applications.

In this thesis, we will model sales time series of some Walmart stores in the United States. We will use the Facebook's Prophet modelling framework. It consists of a flexible Bayesian model which can be customised to fit a wide range of possible time series modelling tasks. We will implement some information criteria for the model and compare it to more ad-hoc methods for model validation.

2 Background

Predictions of sales can be used to minimise waste and ensure sufficient supply of goods. Advertising products can yield greater sales, but not all ads are equally effective. Shifting marketing investments to more impactful marketing channels can increase sales while keeping costs the same. Both of these examples demonstrate scenarios where statistical models can be used to avoid making suboptimal decisions.

Models can be viewed as simplifications of reality. This means that no model never really matches the true data generating processes, but if one can formulate a model that works sufficiently, it might be possible to draw some conclusions from them. Most of the models we are concerned with link together some explanatory variables to the measured data. The parameters of the function are learned by fitting the data to the model. The fitted model can then be used to predict future observations, if we can know the explanatory variables beforehand. The fitted model parameters can likewise give insight on the behaviour of the physical world, assuming that said parameters have a sensible interpretation.

Bayesian models differ from frequentist models in two major ways. Firstly, in Bayesian thinking, parameters are not expressed as point values but distributions. Secondly, the distributions are affected also by the prior beliefs of the distribution. This means that the data is not the only source of information affecting the results. These qualities allow for embedding uncertainty and business knowledge in the models. Highly informative priors can in addition make the models behave more robustly in case the data has some outliers.

Model selection methods can be categorised by the view they take on whether the considered models contain the actual data generating model. This data generating model would represent the actual real-world process by which some events result in the observed data. M-closed view assumes that the model is present in the considered models. M-open view instead attempts to model the data with minimal assumptions and is thus more applicable with real-world data.

? describe multiple approaches for model selection. The most often used methods are information criteria, hold-out predictive, and cross validation. None of the most popular methods make explicit assumptions of the "true" model and are thus rather simple to implement. The methods differ in how they reuse data when evaluating model performance. Information criteria are based on calculating the training utility of the model and adding a bias term that attempts to correct for the reuse of training data in model evaluation to prevent overfitting to the available data. Hold-out predictive methods separate the data to a training and testing set. The testing set is used to evaluate the utility function for the model that has been trained using the training data. The method is considered to be robust if enough training and testing data is available. Cross validation takes this hold-out approach a step further by splitting the dataset and training the model multiple times. This allows for all of the data to be used in testing once. If the amount of available past observations is limited this can help with robustness of evaluation.

3 Datasets and methods

3.1 Original data and data processing

The original dataset consists of department-level weekly sales data of 45 stores. The data spans a little over two years, from 20 February 2010 to 26 October 2012. Each store location is associated with features that may help with sales prediction. The features are temperature, unemployment, price of gas, and consumer price index CPI. Furthermore, the store size and type of store is known. The original dataset contains information about markdowns, but this feature is available only from 11 November 2011 onwards. This renders the feature useless, as the effect of the missing markdowns will not be attributed correctly, skewing the results. For example, a spike in sales could be due to markdown campaign or seasonality. If the markdown data would be complete for the whole dataset the effect could be separated from seasonality. However, as there is markdown data for less than two years, some spikes will not have data about possible markdowns. This leads to the uplift to be attributed to seasonality, which may not be correct. Leaving markdowns out of the features will effectively cause the uplift from markdowns to be attributed to noise if the markdown campaigns are not seasonal.

The original dataset contains sales data on a department level. There are a total of 81 departments, but not all are present in every store. We are still left with 3331 sales time series. Not all time series can be modelled or fitted, thus we need to reduce the amount of data to have any hope of modelling anything in a reasonable time. The department-level sales are first aggregated to store-level sales by summing sales in each department in each store. This still leaves us with 45 time series, which would still require too much time to model each individually. To simplify this issue, the stores are further aggregated based on store type (A, B or C). The aggregated sales are presented in Fig. **??**. The features for these store groups are obtained by calculating a size-weighted average. As the aggregation process of the features can render the features less meaningful, a small subset of stores is also selected from each store group. Comparing these stores to the store groups will reveal if the aggregation lessens the predictive performance of the models using the provided features. The aggregated features can be found in Fig. **??**.

The sales data for store groups A and B demonstrate strong seasonality. Especially during Decembers, the turnover spikes due to strong holiday effect. For store group C the sales are more consistent throughout the year. Temperature is the only feature that clearly behaves seasonally. Originally, unemployment was expressed to equal the latest unemployment statistic. This made the feature behave rather discontinuously. To get around the discontinuity affecting the results, the unemployment for each week was calculated by interpolating unemployment from last and next changepoint. After the transformation, both unemployment and consumer price index (CPI) have a clear trend.

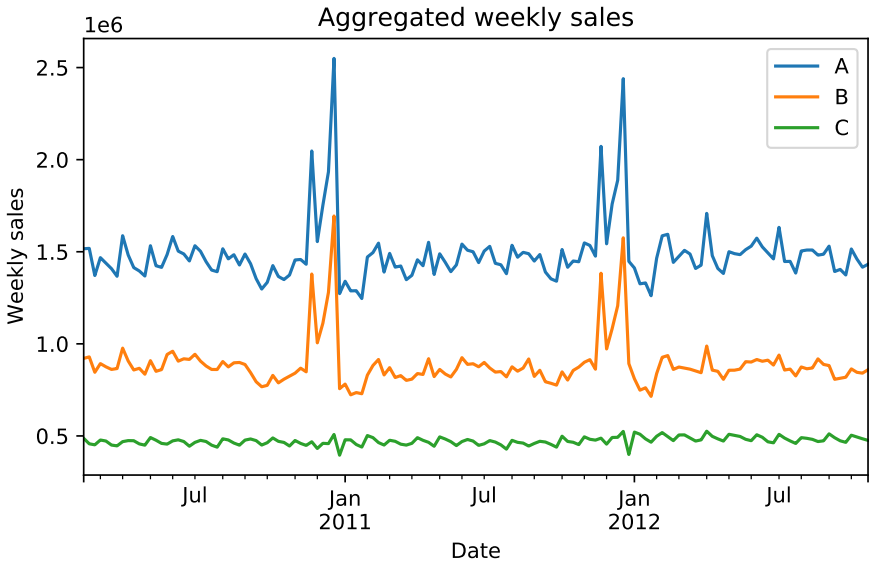
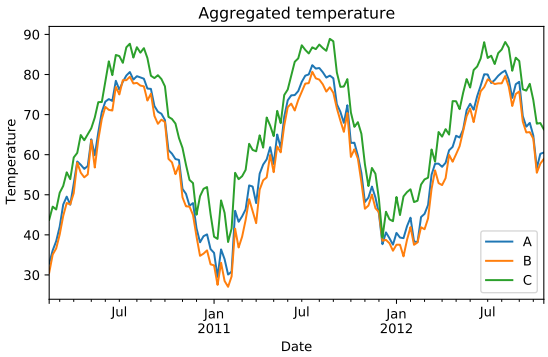


Figure 1: Aggregated weekly sales by store group. Groups A and B demonstrate holiday effects while group C remains relatively stable through each year.

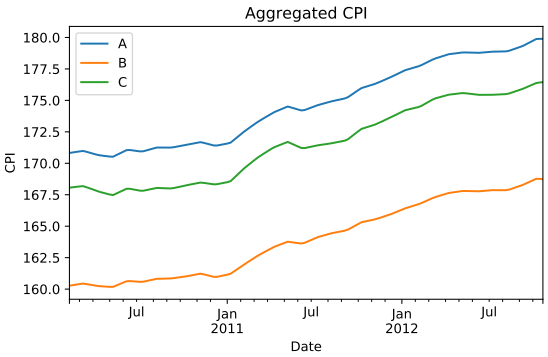
3.2 Prophet-models

The Prophet model by **?** is a linear model that is used to predict time series. The model consists of seasonality components, a piece-wise linear trend, and possible explanatory variables such as holidays. A component-wise breakdown of a fitted model is visualised in Fig. **??**. The seasonalities are modelled using a finite Fourier series approximation. Trend is modelled linearly between several pre-determined changepoints. After the last change point the model assumes a linear trend. The prophet model treats time as a feature ranging from 0 to 1. Notably, the prediction does not directly depend on previous observations but seasonality and trend. They are calculated also using the other observations, but that calculation does not require a constant time difference between observations. As such, the model can handle inconsistent time intervals between datapoints. This means that the training and testing set do not have to consist of sequential observations.

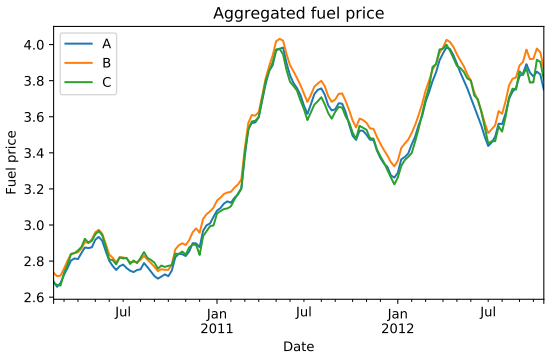
Each of the five used models is a Prophet model with different settings and features. The models are labelled M1 to M5. The first model was chosen to act as a benchmark to be compared against the other models. This should demonstrate if changes to the base model yield any improved predictive performance. The second model adds on the benchmark model by also incorporating the additional features present in the dataset. These were temperature, CPI, unemployment rate and fuel price. The fourth model M4 does not utilise the extra features. It attempts to demonstrate the behaviour of the metrics with overfitted models. The number of Fourier coefficients for seasonality are doubled from the benchmark model. M3 and M5 differ fundamentally from the other models as they are not reasonably possible to implement in a real-world prediction scenario. Both of them are given the sales time series as an explanatory variable. This is done to study the behaviour of the metrics



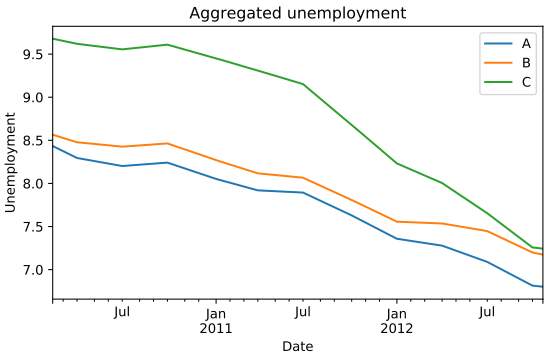
(a) Temperature by store group



(b) Consumer price index by store group



(c) Fuel price by store group



(d) Interpolated unemployment rate by store group

Figure 2: Additional features used in Model M2.

Table 1: Different settings used in each model. Extra features are fuel price, CPI, unemployment and temperature.

	M1	**M2**	**M3**	**M4**	**M5**
Yearly seasonality coefficients	10	10	10	20	None
Extra features		X			
Trend changepoints	25	25	25	25	1
Response variable as regressor			X		X

in situations where the models actually represent the real-world data generating process. In this case the process is rather simple, as the correct prediction should follow trivially from the features. M3 is equal to the benchmark model extended by this feature. M5 has also less trend changepoints and disabled seasonality. As M5 should produce the exact same output with less variables the information criteria should always rank it as better performing. The differences between used models are displayed in Table **??**.

3.3 Metrics

The models were compared using five different metrics. The metrics used were Akaike information criterion (AIC), deviance information criterion (DIC), Watanabe-Akaike information criterion (WAIC), mean average percentage error (MAPE), and 10-fold CV. The information criterions are calculated using only in-sample fits. They aim to estimate the out-of-sample predictive performance of the model using the likelihood of past observations. Notably, only WAIC of the information criterions fully considers the Bayesian nature of the model. It utilises the whole posterior distribution instead of some aggregated values. MAPE is calculated in the spirit of hold-out predictive methods by reserving a small portion of data from the tail of the dataset to act as a test set. The percentage errors of the predictions are then averaged to obtain MAPE. 10-fold cross validation is calculated by assigning randomly 1/10 of the datapoints as test set and then calculating the likelihood of the prediction for the test set. This is performed ten times so that each datapoint belongs to the test set once. 10-fold CV is the average of the results. Test set selection of MAPE is demonstrated in Fig. **??** and one of the 10-fold CV splits can be found in Fig. **??**.

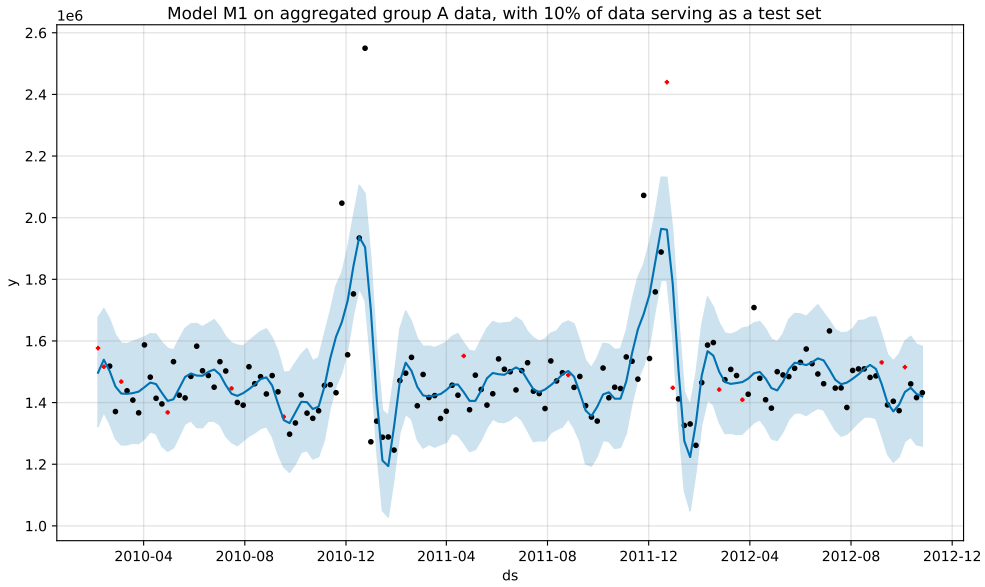


Figure 3: Demonstrating the CV metric calculation. Red dots represent the 10% of the datapoints that were selected for testing set. Model is fitted using training data represented by the black dots. This data splitting and evaluation is repeated ten times until all datapoints have belonged to the test set exactly once.

3.4 Bayesian model fitting

Bayesian models rely on inferring the posterior probability densities of the parameters from data and prior probabilities. For complex models, this cannot be done analytically but by using Markov chain Monte Carlo sampling (MCMC). Stan (**?**) is

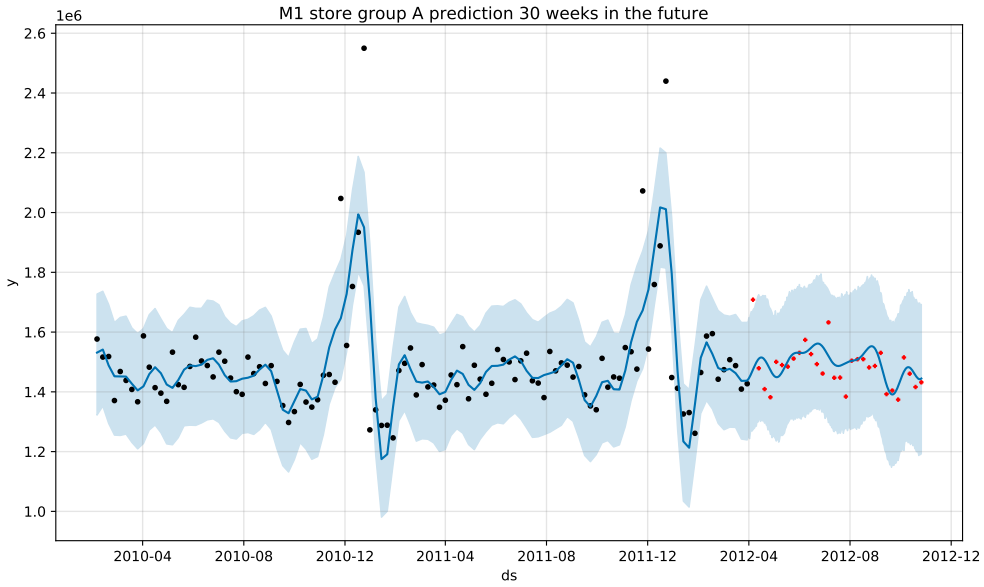


Figure 4: Hold-out prediction 30 weeks in the future. Training data in black, test data in red. This train / test split is used in MAPE metric.

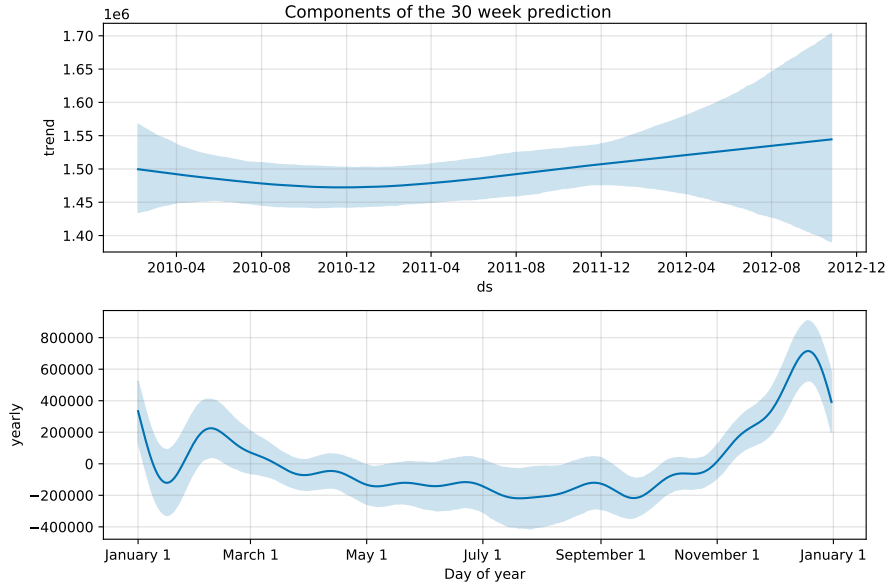


Figure 5: Trend and yearly seasonality components of the fitted model used in Fig. ??.

a Bayesian data modelling platform. In this thesis Stan is used to mathematically model the Prophet model and fit it using MCMC sampling. To be more specific, the adaptation of MCMC used in Stan was the No-U-Turn sampler (NUTS), as described in **?**. This sampling process explores the possible parameter space of the model, searching for likely parameter values, and yields a large collection of point value combinations for the parameters. Usually, the sampling process utilises multiple

chains to ensure proper convergence of the sampling process.

3.5 Workflow

This thesis focuses on the metrics used in model selection, not necessarily the model selection process itself. The workflow of metric evaluation was as follows: First the data was analysed to check for which features to use or if some data processing had to be done. After this the data was aggregated to different datasets: three by store group level and nine store level datasets. The compared models were constructed from the available data. Each of the models was fitted for each of the available datasets. Each of the five metrics were then computed for all the model-data combinations. All the fitted models and metrics were saved in a Pickle container to be analysed later. Ordering of the models by metrics results were then compared against each other within each dataset.

4 Results

The models were fitted using Stan for MCMC sampling with four chains and sampling parameters target metropolis acceptance rate $\delta = 0.9$ and maximum tree depth of 11. Each chain was sampled for a total of 1 000 times of which 500 were discarded as warmup samples. The other sampling parameters were left to their default values as suggested by **?**. The obtained results for the store group level metrics are presented in Table **??**. Results for store level data are presented in the Appendix **??**. The columns of the tables represent the evaluated metrics. For the information criteria and MAPE evaluation lower values are desirable. As 10-fold CV is a likelihood, higher is better. The metric results can be used to rank the models by their performance. This is visualised in Fig. **??**. From the figure it can be seen that the results are somewhat consistent between the models and metrics.

The visual representation highlights the order different metrics rank the models. AIC and DIC ranked all the models similarly for all the datasets, but there were differences in the margins between the models. WAIC seemed to mostly agree with AIC and DIC, but in datasets B and C the order of models M1 and M4 was reversed.

Ranking by MAPE has many similarities with the information criterion rankings. However, the Model M5 performed worse with all datasets, dropping behind M3 in A and C, but also behind M1 and M4 in B. In dataset B MAPE ranked M1 better than M4. This agrees with the WAIC ranking, but it is the other way around with AIC and DIC. The other notable difference in MAPE scores is that in datasets A and B the Model M2 seems to perform much worse than the other models. However, this phenomenon is not present in C.

The 10-fold cross validation ranks the models M5 and M3 the best in all datasets, just like the information criterions do. However, the ranking of the rest varies. In dataset C the models M1, M2 and M4 are ranked similarly as by AIC and DIC. In both A and B, the Model M4 takes the last place, while M2 and M1 get a score similar to each other. M2 was ranked better than M1 for dataset A and vice versa

for B, but the margin was rather slim.

4.1 Applicability of the metrics

Overall, it seems that the metrics correctly recognise the "data generating models" from other models. Additionally, M5 mostly ranks better than M3, as it has less non-informative regressors. When it comes to the ranking of models M1, M2 and M4, the results are not as clear. In store groups A and B, M2 performs much worse than the other models when evaluating with MAPE. This likely indicates, that the model overestimates the importance of the external regressors. As the regressors have trends, the prediction drifts off the actual sales. In the 10-fold CV this is not observed, as the datapoints in test set are picked randomly. The test points have similar-valued training points next to them, thus the trend does not have time to cause the predictions to drift off.

Model fitting and metric evaluation times can be found in Table **??**. Fitting and evaluation was performed using two cores on a 2017 MacBook Pro with an 2.3GHz Intel Core i5-7360U CPU and 8GB of RAM. For AIC and WAIC the times are averages of hundred evaluations, for DIC average of three evaluations, and rest are single runs. Fitting time depended heavily on the complexity of the model. The metrics can be categorised by their evaluation time. Both AIC and WAIC took under 10ms for all models. DIC evaluation took between six to nine seconds for all models, but this does not seem to depend on model complexity or initial fitting time. Lastly there are both MAPE and 10-fold CV evaluation. Their evaluation times did indirectly depend on model complexity, as both metrics require the model to be refit with different splits of data. Their evaluation took between 2 to 4.5 and 7 to 11 times their respective model fitting time. The fastest category of AIC and WAIC are calculated by using only matrix operations on the sampled parameters. This makes their evaluation extremely rapid. DIC also relies only on the fitted parameters, but first maximum likelihood estimates are calculated using the prediction method of the model. This involves non-matrix operations that are considerably slower to execute.

The metrics were tested for stability and consistency by plotting the metric value as a function of used samples. First model M5 was fitted with 500 warmup iterations and 4000 samples per chain. The metrics were calculated by giving the evaluation function slices of the sampled parameters corresponding to the evaluated sample count. For AIC, WAIC and 10-fold CV metrics were evaluated every 10th sample, for DIC and MAPE every 100 samples. These resolutions were chosen to better manage the computation time, as both DIC and MAPE relied on predictions using sampled parameters. The results for metrics in devinace scale are presented in Fig. **??** and for MAPE in **??**. Notably, for metric in deviance scale the level of noise after 500 samples is low, only under one unit of deviance. MAPE also behaves well with changes in metric value of under 0.01% after 500 samples. For all metrics the observed differences in model performances from Fig. **??** greatly exceed the level of noise in metric values.

When examining the results, it is important to keep in mind the different utility functions used in the metrics. Other metrics than MAPE are based on estimating

Table 2: Numerical representation of the validation metric results applied by store group.

(a) Store group A

Model	AIC	DIC	WAIC	MAPE	10-fold CV
m1	-394.4	-438.8	-15.9	0.0336	205.4
m2	-384.6	-431.5	136.5	0.0445	208.2
m3	-654.5	-702.1	-669.3	0.0219	349.6
m4	-457.4	-512.4	-41.6	0.0275	164.1
m5	-703.9	-703.2	-707.8	0.0253	351.3

(b) Store group B

Model	AIC	DIC	WAIC	MAPE	10-fold CV
m1	-374.3	-419.1	19.3	0.0279	204.2
m2	-365.5	-412.2	315.9	0.0494	196.9
m3	-647.8	-695.3	-663.2	0.0255	347.3
m4	-446.9	-501.4	152.2	0.0286	162.1
m5	-699.4	-698.3	-701.5	0.0297	348.9

(c) Store group C

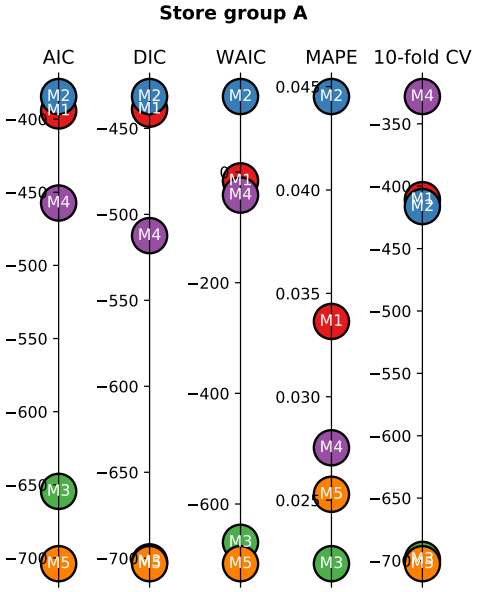
Model	AIC	DIC	WAIC	MAPE	10-fold CV
m1	-454.0	-501.2	-397.1	0.0339	247.4
m2	-447.1	-496.6	-186.4	0.0349	244.2
m3	-668.9	-715.6	-668.1	0.0136	357.0
m4	-549.1	-605.1	-214.3	0.0253	291.4
m5	-741.9	-740.6	-743.7	0.014	370.2

the log-likelihood of out-of-sample predictions, while MAPE evaluates the mean absolute percentage error. Log-likelihood is proportional to MSE as the model consist of a normally distributed error term with a constant variance. In a perfect world the utility function would represent the actual losses due to errors in predictions. However, it is difficult to estimate those accurately.

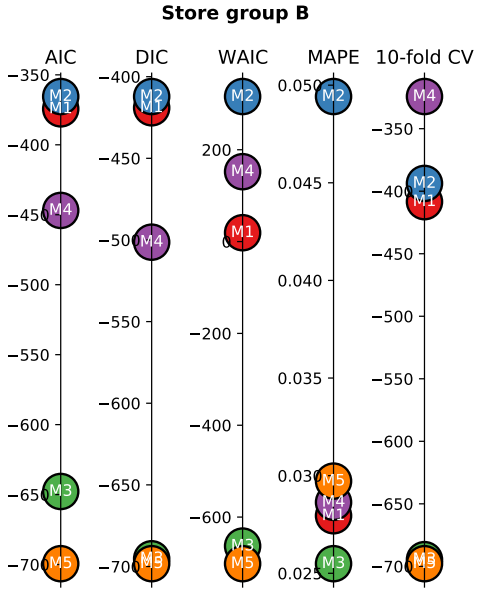
From a business application point of view, it can be concluded that the 10-fold CV is not a feasible model selection metric. In practice the used models are very complex and may take hours to fit. This renders cross validation-based methods useless as the results are often needed rapidly. If the dataset has enough datapoints, hold-out based metrics like MAPE and others with possibly different utility functions may be used. In a scenario with limited data availability the information criteria may suffice.

5 Summary

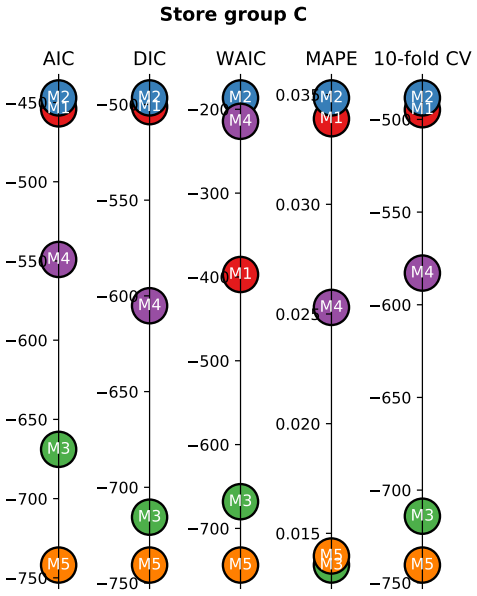
In this thesis we studied the applicability of multiple model selection metrics using datasets and Bayesian models typical to business settings. The studied metrics demonstrated consistent results in ranking model performance, making them useful



(a) Store group A



(b) Store group B



(c) Store group C

Figure 6: Visualisation of model validation metric results by store group. Arranged such that visually lower is better. Each axis is scaled so that extreme values are at the ends. Please note the inverted axis of 10-fold cross validation metric. For precise values please refer to Table **??**

Table 3: Model fitting and metric evaluation times in seconds.

	Initial fitting time	AIC	DIC	WAIC	MAPE	10-fold-CV
M1	32.27	0.00207	8.31	0.00420	108.39	358.46
M2	59.47	0.00326	7.88	0.00412	163.33	491.66
M3	105.55	0.00740	8.06	0.00453	203.68	763.72
M4	65.30	0.00312	7.48	0.00426	151.36	490.07
M5	7.68	0.00376	6.69	0.00453	34.79	83.10

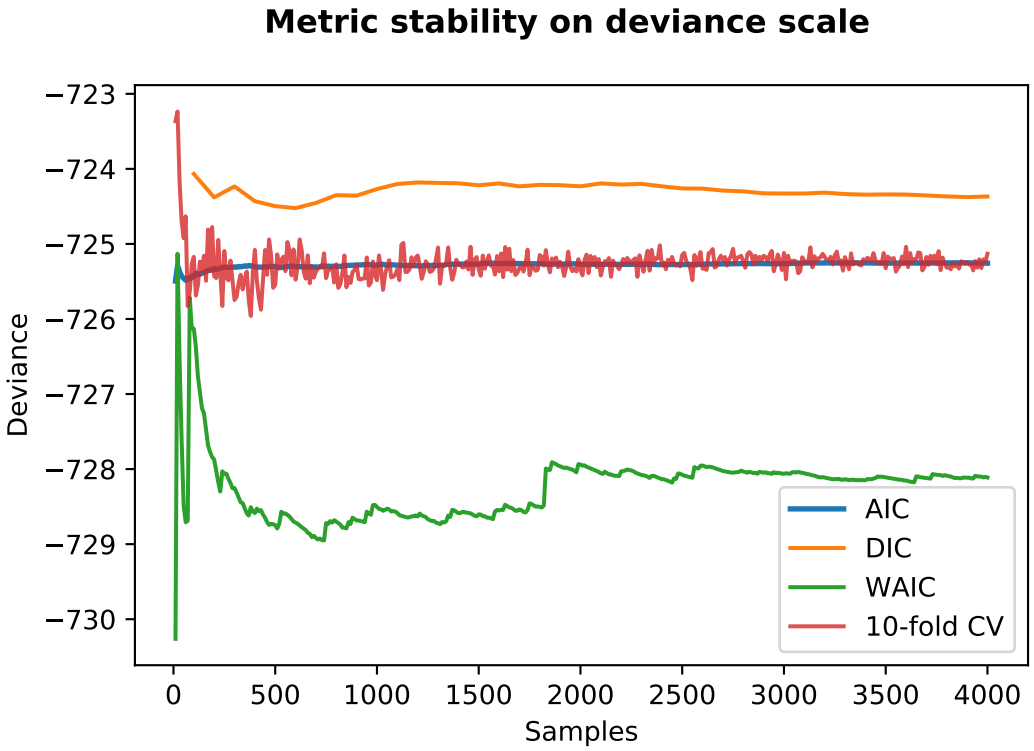


Figure 7: Metric values by sample count. Lower noise indicates more stable results. Metrics in deviance scale are present here, for MAPE see Fig. **??**

for model selection. The short time interval was not an issue for all metrics, as the in-sample-metrics were not suffering similarly from the lack of data.

To study further the applicability of common model selection metrics one could experiment with a retail dataset with better data quality. The effect of marketing and discounts could greatly affect sales and thus make for an interesting modelling problem. Longer time series would be desirable, as it would help with separating seasonality from the effect of explanatory variables.

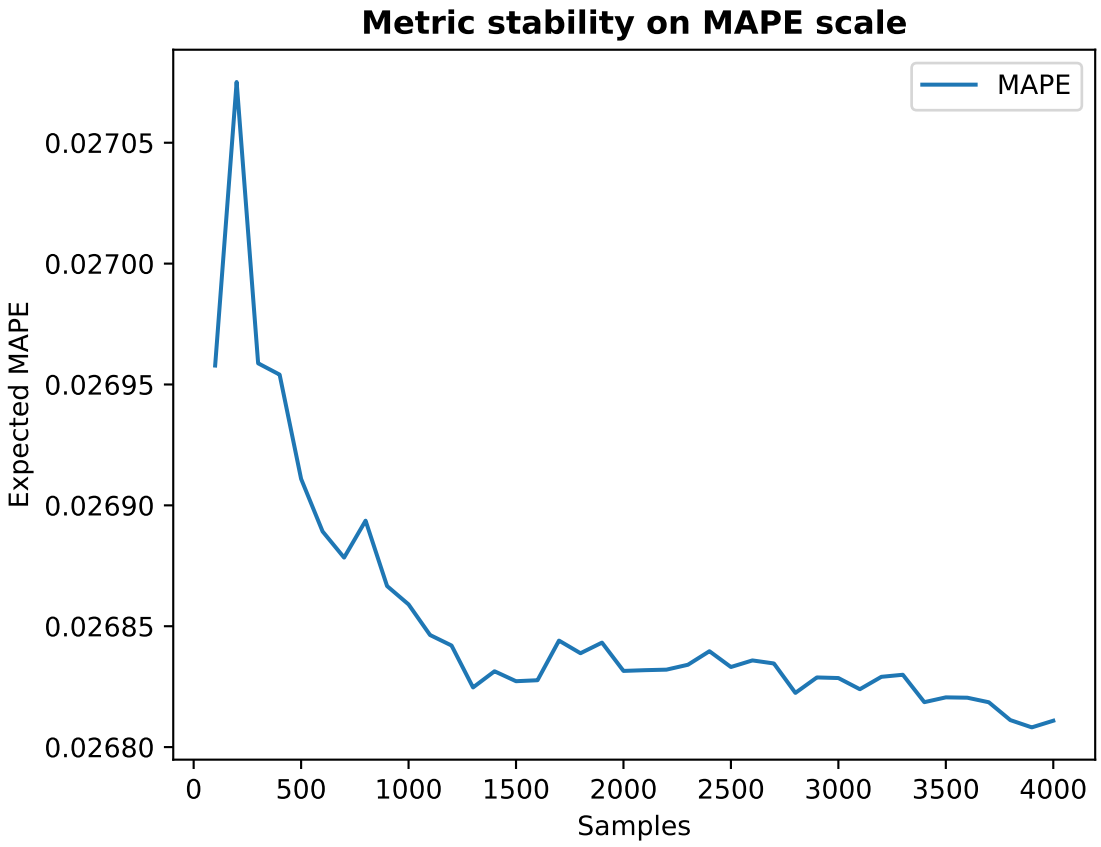


Figure 8: Expected MAPE by sample count. Lower noise indicates more stable results. Metrics in deviance scale are presented in Fig. **??**

A Results for store-level data

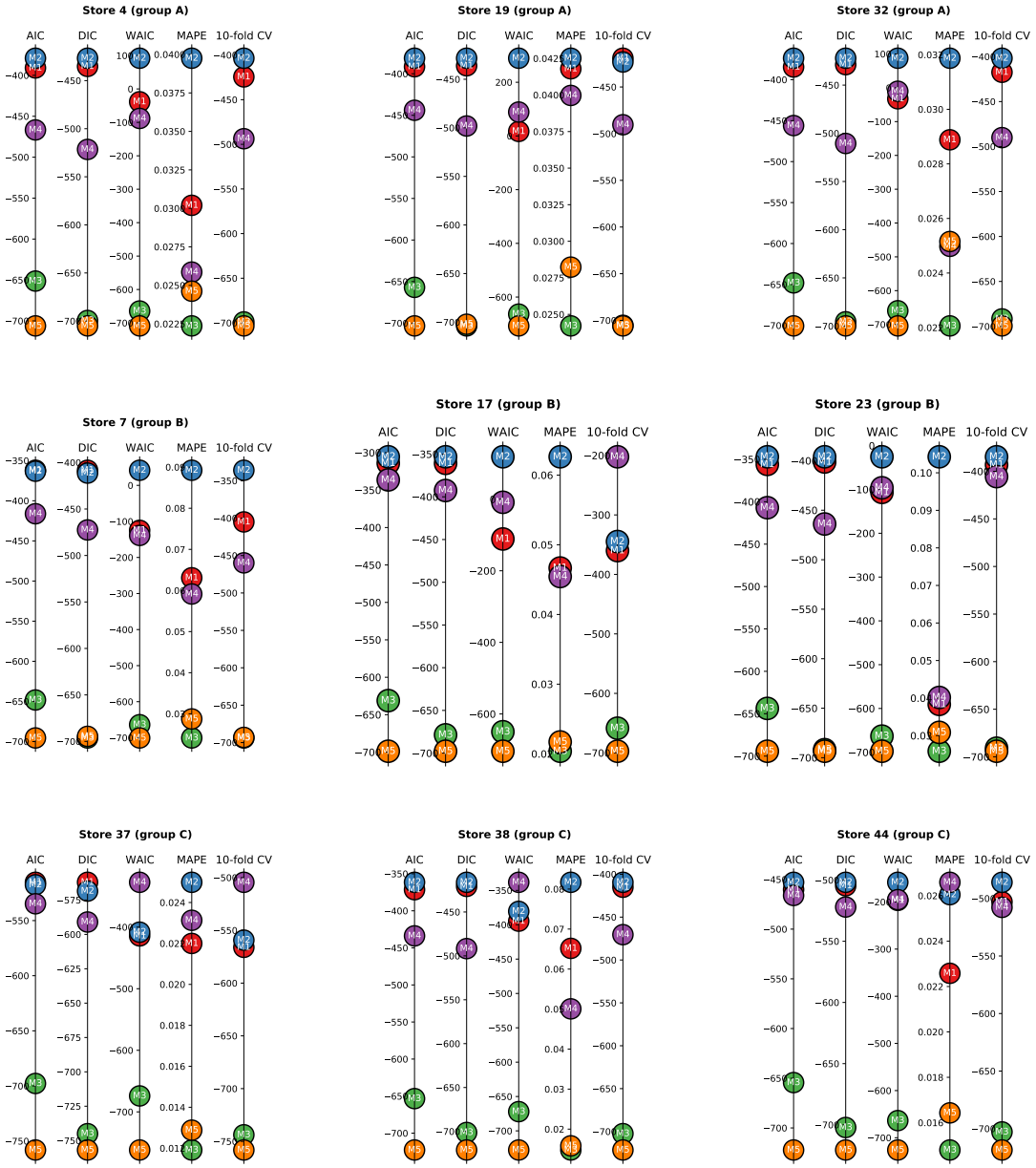


Figure A1: Visualisation of model validation metric results for single stores.