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** 13+2 **Language** English

Abstract

Your abstract in English. Keep the abstract short. The abstract explains your research topic, the methods you have used, and the results you obtained.

Keywords Bayesian models, model validation, information criterion, Facebook Prophet

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 TkT Mikko Ervasti, TkT Paavo Niskala

Päivämäärä 14.10.2020 **Sivumäärä** 13+2 **Kieli** Englanti

Tiivistelmä

Tiivistelmässä on lyhyt selvitys kirjoituksen tärkeimmästä sisällöstä: mitä ja miten on tutkittu, sekä mitä tuloksia on saatu.

Tämän opinnäytteen tiivistelmäteksti kirjoitetaan opinnäytteen luettavan osan lomakkeen lisäksi myös pdf-tiedoston metadataan \thesisabstract-makron avulla (kasto yllä). Kirjoita tähän luettavaan tiivistelmälomakkeeseen menevä teksti. Tässä saa olla erikoismerkkejä kuten kreikkalaiset kirjaimet ja rivinvaiho- ja kappaleenjakomerkit. Tämän tekstin on muuten oltava sama kuin metedatatiivistelmän teksti.

Jos tiivistelmäsi ei sisällä erikoimerkkejä eikä kaipaa kappaleenjakoa, voit hyödynttää makroa \abstracttext luodessasi lomakkeen tiivistelmää (katso kommentti alla).

Avainsanat Bayeslaiset mallit, mallin validointi, informaatiokriteeri, Facebook Prophet

Contents

1 Introduction

Businesses need to forecast a multitude of things to succeed. As companies collect more and more data the possibilities of forecastable subjects and possible features increase dramatically. However, one can't just thow more features at a model and except it to perform well. Also not all models are suitable for all forecasting tasks. The amount of possible models is endless, but most of them are bad. To find the usefull ones we must be able to measure the goodness of a model.

Model validation is an integral part of a robust modelling framework. Bayesian models are a relatively novel model type which has gained a lot of populatiry 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 criterions which try to predict out of sample fit can have strong biases and some may not take into consideration the nature of bayesian models and distributions of parameters. Cross validation can be computationally extremely intensive. Also most information criterions are based on a loss function proportional to the root mean square of errors, but that may not always be the most usefull function to optimize for. In some businesses the consecuences of errors in forecasts can be described better with the mean absotule percentage error. All in all there does not exist a clearly better method for solving all bayesian model validation problems.

In this paper we will model sales timeseries of some Wallmart stores in the United States. We will be using the Facebook's Prophet modelling framework. It consists of a flexible bayesian model which can be customized to fit a wide range of possible time series modelling tasks. We will implement some information criterions for the model and compare it to more ad-hoc methods for model validation.

2 Background

In retail business many things can be a subject for optimizing. Predictions of sales can be used to minimize 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 impactfull 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 unoptimal decisions my manual guesswork.

Models can be viewed as simplifications of reality. This means that no model never really mathces the true data generating processes, but if we can formulate a model that works well enough we can try 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 also

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. In bayesian thinking parameters are not expressed as point values but distibutions. Uncertainty can be expressed with wider distributions. New data is not the sole source of the newly fitted parameters as it is merely used to update prior beliefs about parameter distribution.

2.1 Rakenne

3 Datasets and methods

The original dataset consists of department-level weekly sales data of 45 stores. The data spans a little over two years, from 20.2.2010 to 26.10.2012. 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. Also the store size and type of store is known. The original dataset contains information about markdowns, but this feature is available only from 11.11.2011 onwards. This renders the feature useless, as it is hard to compensate for the missing data.

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 seem a lot more consistent. 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 dincontinously. 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.

The original dataset contains sales data on a departement level. There are a total of 81 departements, but not all are present in every store. We are still left with 3331 sales timeseries. Not all timeseries can be modelled of fitted, thus we need to reduce the amount of data to have any hope of modelling anything in a reasonable time. The departement-level sales are first aggregated to store-level sales by summing sales in each departement in each store. This still leaves us with 45 timeseries, which is still a bit too much. To simplify things a bit more, the stores are further aggregated based on store type (A, B or C). 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 meaningfull, a small number of stores is also selected for modelling from each store group.

The prophet model is a linear model used to predict time series. The model consists of seasonality components, trend, and possible explanatory variables. The seasonalities are modelled using a fourier approximation with a limited amount of coefficients. Trend is modelled linearly between a number of pre-determined changepoints. After the last change point the model also assumes a linear trend. The

prophet model handles time as a feature ranging from 0 to 1. As such, the actual time interval between datapoints does not have to be constant.

Each of the five fitted models is an prophet model with different settings and features. The models were constructed to showcase the behaviour of the metrics in situations where the model is used as one could use it in a busiess environment, overfitting seasonality, a case where a data-generating model can be constructed from the explanatory variables, and a case similar to the last-mentioned but with even less distracting variables. The models were as follows:

M1: Only weekly sales, no additional features. M2: All features, no other trics. M3:

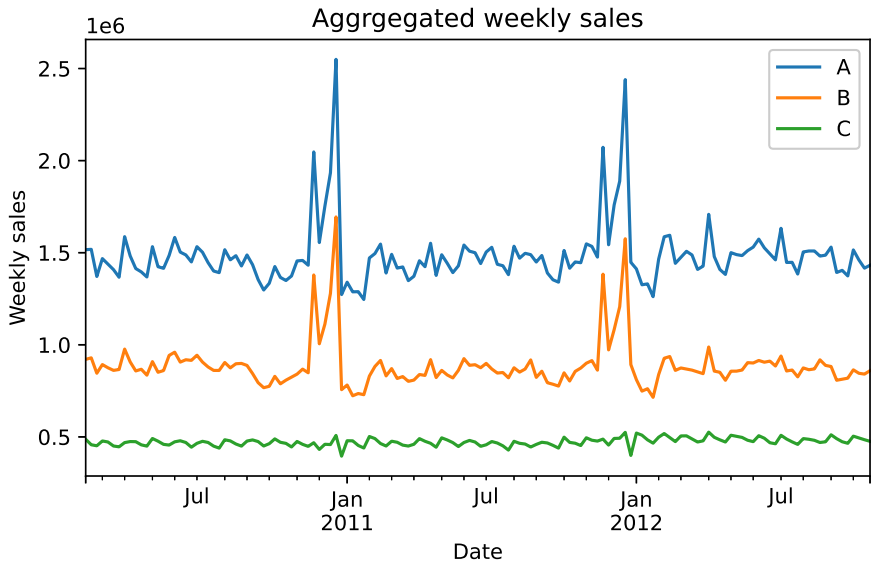
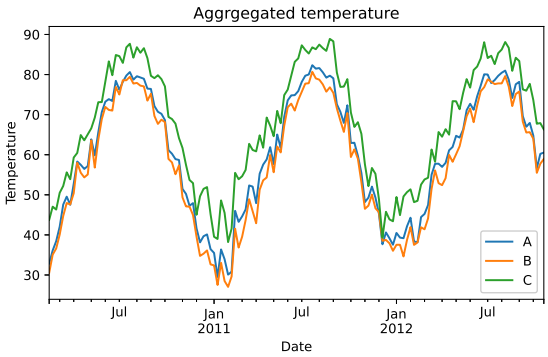


Figure 1: Aggregated weekly sales by store group. Groups A and B demonstrate holiday effects while group C remains relatively stable through each year.

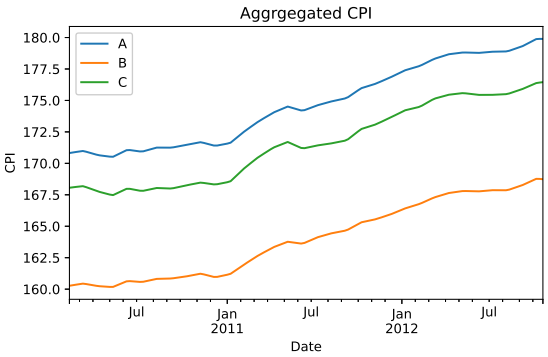
4 Results

Applying the metrics yielded promising results. The results are presented visually in figures 3, 4 and 5. 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.

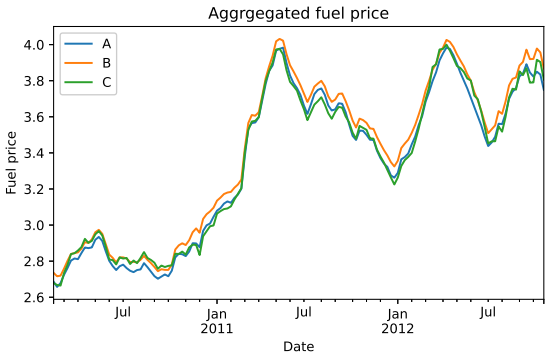
Rankign by MAPE has many similarities with the information criterion rankings. However the model m5 performed worse with all datasets, dropping behind m3 in A and C, but also behind m1 and m4 in B. In dataset B MAPE ranked m1 better than m4. This agrees with the WAIC ranking, but it is the other way around with AIC and DIC. The other notable difference in MAPE scores is that in datasets A



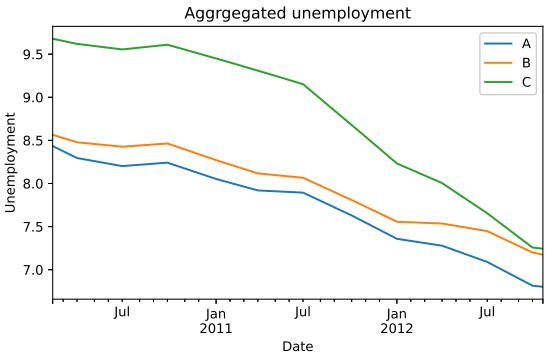
(a) Temperature by store group



(b) Consumer price index by store group



(c) Fuel price by store group



(d) Interpolated unemployment rate by store group

and B the model m2 seems to perform much worse than the other models. However this phenomenon is not present in C.

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.

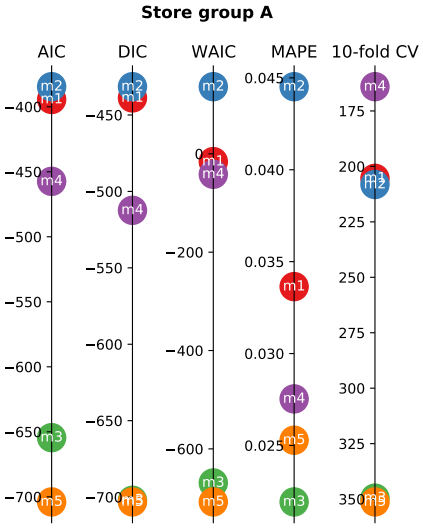


Figure 3: Visualisation of model validation metric results applied on store group A. Arranged such that visually lower is better. Each axis is scaled so that extreme values are at the ends. Please note the inverted axis of 10-fold cross validation metric. For precise values please refer to table 1.

Table 1: Results of validation metrics applied on 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

5 Summary

References

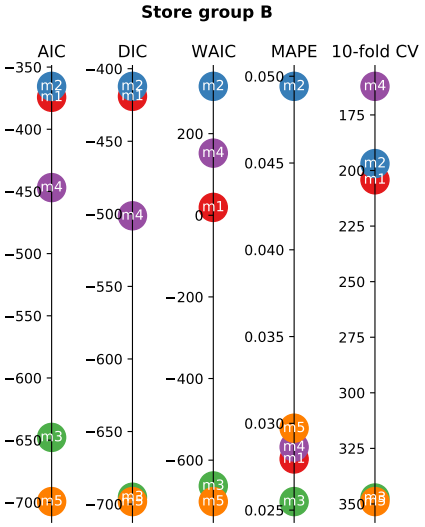


Figure 4: Caption this

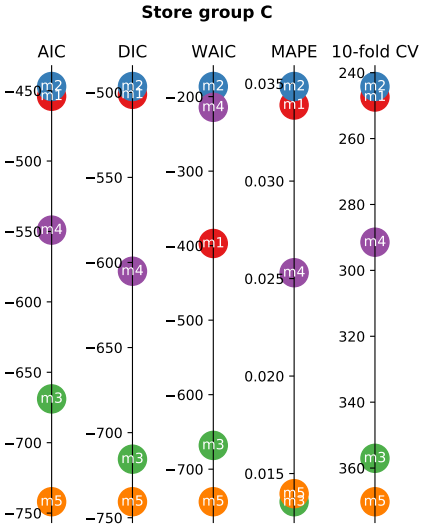


Figure 5: Caption this

Table 2: Results of validation metrics applied on 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

Table 3: Results of validation metrics applied on 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

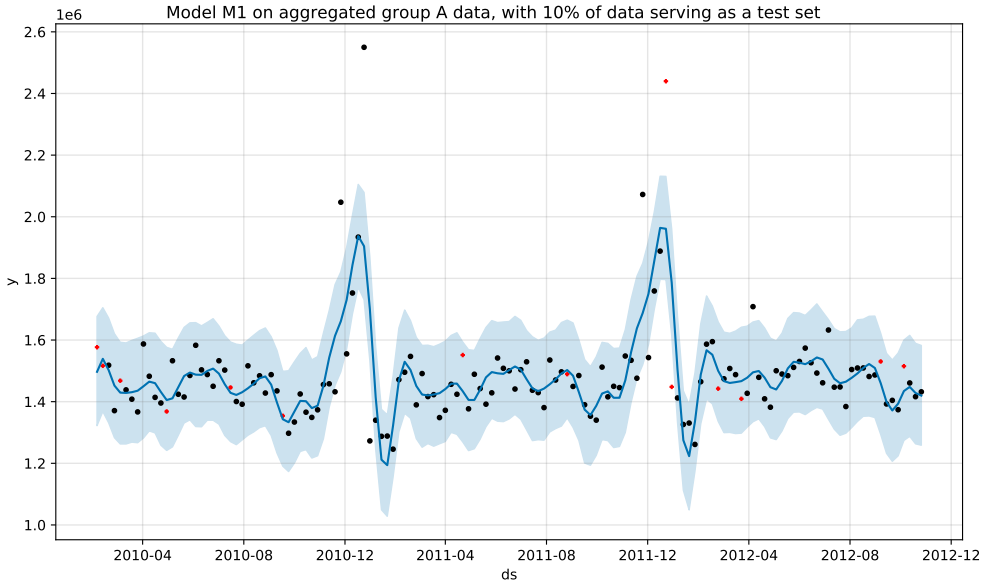


Figure 6: Demonstrating the cv metric calculation

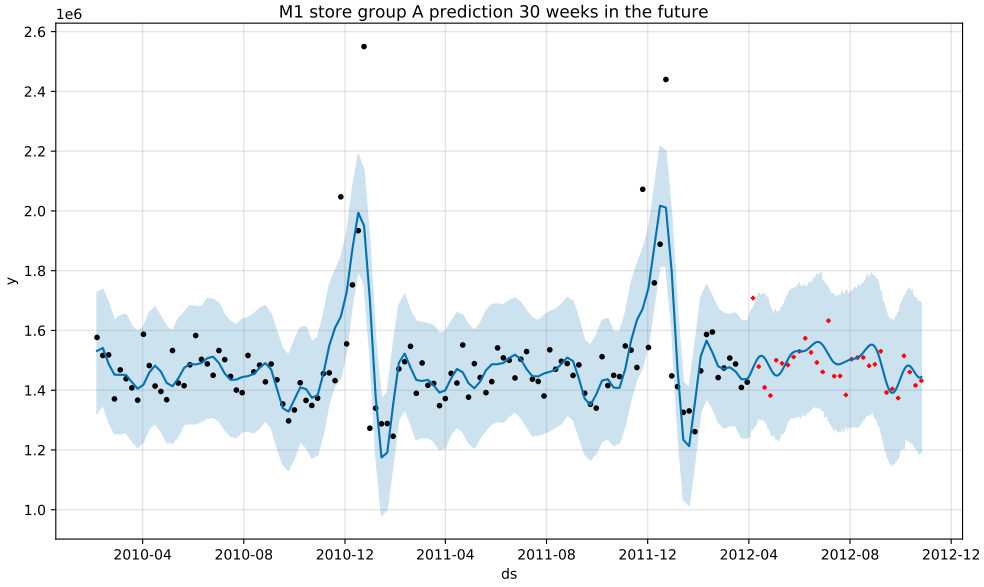


Figure 7: A prediction 30 weeks in the future. Training data in black, test data in red.

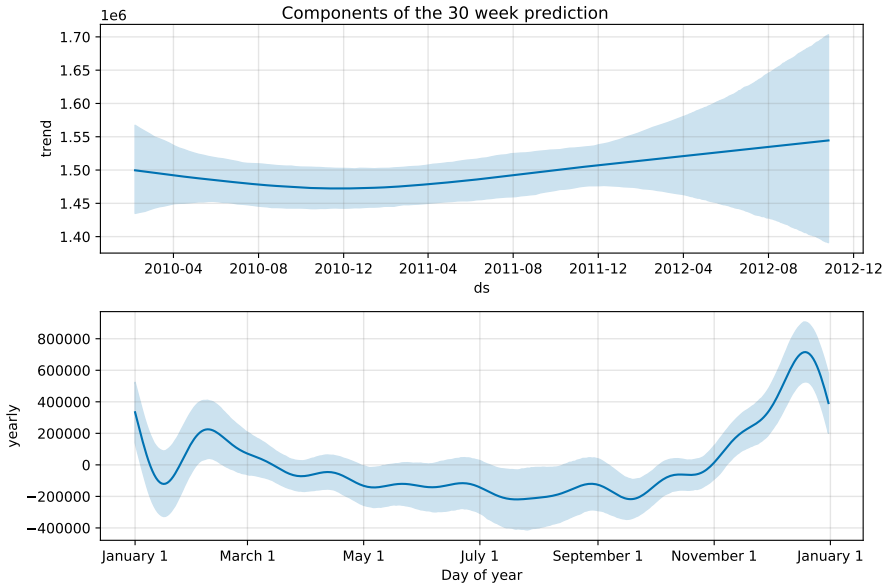


Figure 8: Components of the fitted model used in figure 7

A Esimerkki liitteestä

B Toinen esimerkki liitteestä