Bachelor's Programme in Science and Technology

Projection Predictive Method for Bayesian Model Selection in Retail Time Series Forecasting

Elias Ylä-Jarkko

Aalto University

BACHELOR'S THESIS

Aalto University
BACHELOR'S THESIS 2022

Projection Predictive Method for Bayesian Model Selection in Retail Time Series Forecasting

Elias Ylä-Jarkko

Thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Technology.
Otaniemi, 21 Jan 2022

Supervisor: professor Fabricio Oliveira
Advisor: FM Ella Tamir

Aalto University
School of Science
Bachelor's Programme in Science and Technology

Author
Elias Ylä-Jarkko

Title
Projection Predictive Method for Bayesian Model Selection in Retail Time Series Forecasting

School School of Science

Degree programme Bachelor's Programme in Science and Technology

Major Mathematics and Systems Analysis **Code** SCI3029

Supervisor professor Fabricio Oliveira

Advisor FM Ella Tamir

Level Bachelor's thesis **Date** 21 Jan 2022 **Pages** 21+1 **Language** English

Abstract

Forecasting of retail time series is an attractive topic due to the possible cost savings it provides. Bayesian methods are often useful for the task since they provide a more intuitive expression of uncertainty in comparison to traditional frequentist methods. However, there are numerous ways to formulate a Bayesian forecasting model, which is why it is important to craft the best possible model case by case in order to achieve the greatest benefits. Models can be improved with variable selection methods, which aim to simplify models and thereby achieve greater forecast accuracy.

In this thesis we model the daily total sales of an Ecuadorian grocery retailer with a Bayesian time series model and perform variable selection using the projection predictive method. Namely, variables are selected by fitting models with some explanatory variables to the predictions of a model that includes all the available explanatory variables. We find the relative order of importance of the variables by comparing the predictive performance of models with different subsets of variables. Furthermore, we are able to determine the smallest subset of explanatory variables that constitute to a model that has similar predictive performance as the model with all explanatory variables included.

The results indicate that the daily total sales of the retailer can be predicted with satisfactory accuracy using only one out of the eight available explanatory variables, assuming the best variable is chosen. We find that from the available explanatory variables, which include weather data, economic indicators and promotion information, the most significant one is the price of potatoes. Including further explanatory variables did not improve model performance significantly. We found that the projection predictive method is applicable to selecting variables in retail time series models.

Keywords Bayesian, projpred, retail, projection predictive, Prophet

urn https://aaltodoc.aalto.fi

Tekijä
Elias Ylä-Jarkko

Työn nimi
Projection Predictive Method for Bayesian Model Selection in Retail Time Series Forecasting

Korkeakoulu Perustieteiden korkeakoulu

Koulutusohjelma Teknistieteellinen kandidaattiohjelma

Pääaine Matematiikka ja systeemitieteet **Koodi** SCI3029

Vastuuopettaja professori Fabricio Oliveira

Ohjaaja FM Ella Tamir

Työn laji Kandidaatintyö **Päiväys** 21.1.2022 **Sivuja** 21+1 **Kieli** englanti

Tiivistelmä

Vähittäiskaupan myynnin ennustaminen on houkuttelevaa, sillä osuvien ennusteiden avulla voidaan alentaa liiketoiminnan kustannuksia. Erityisesti bayesilaiset menetelmät ovat hyviä ennustamiseen, koska ne kuvaavat epävarmuutta helppotajuisemmin kuin perinteiset frekventistiset menetelmät. Ennuste koostuu tällöin todennäköisyysjakaumista, jolloin on helppoa arvioida erilaisten myyntimäärien toteutumisen todennäköisyyttä. Ennustavia bayesilaisia aikasarjamalleja voidaan silti luoda lukemattomia erilaisia, joten suurimman hyödyn saavuttamiseksi on tärkeää luoda mahdollisimman hyvä malli tapauskohtaisesti. Bayesilaisia malleja voidaan parannella muun muassa muuttujanvalintamenelmin, joiden tarkoituksena on yksinkertaistaa malleja ja siten parantaa niiden ennusteiden laatua.

Tässä työssä mallinnetaan Ecuadorissa toimivan vähittäiskaupan kokonaismyyntiä bayesilaisella aikasarjamallilla ja valitaan sen muuttujia projektioprediktiivisellä menetelmällä. Kyseisessä menetelmässä kerätään ensin ennuste käyttämällä mallia, joka hyödyntää kaikkia saatavilla olevia selittäviä muuttujia. Niitä on tässä työssä yhteensä kahdeksan. Sen jälkeen aiemmin mainittuun ennusteeseen sovitetaan osan muuttujista sisältäviä malleja. Muuttujien tärkeysjärjestys selvitetään vertailemalla mallien ennustuskykyä. Lisäksi etsitään pienin mahdollinen osajoukko muuttujia, joista muodostetulla mallilla on tyydyttävä ennustuskyky.

Työn tulokset osoittavat, että tarkastellun vähittäiskaupan myyntiä voidaan ennustaa tyydyttävällä tarkkuudella jo yhden muuttujan sisältävällä mallilla, kun valitaan muuttujista paras. Tulosten mukaan testatuista muuttujista, joihin kuului muun muassa säätietoja, yleisiä hintaindeksejä ja kaupan alennustietoja, paras selittäjä oli perunoiden hinta. Muiden muuttujien lisääminen malliin ei enää parantanut mallin suorituskykyä merkittävästi, kun perunoiden hinta oli mukana. Projektioprediktiivinen menetelmä osoittautuu soveltuvaksi vähittäiskaupan ennustemallien muuttujien valitsemiseen.

Avainsanat Bayes, projpred, vähittäiskauppa, projektio, referenssimalli, Prophet

urn https://aaltodoc.aalto.fi

Contents

Abstract ii

Tiivistelmä iii

Contents iv

1. Introduction 1

2. Background 3

3. Methodology 6

3.1 Data set 6

3.2 Facebook Prophet model 8

3.3 Projection Predictive Framework 10

4. Results 12

4.1 Order of importance for the explanatory variables 12

4.2 Selecting the optimal submodel 15

5. Conclusions 18

Bibliography 20

A. Appendix 22

1. Introduction

Forecasting demand and sales has been an attractive area of improvement in retail because it provides business benefits, e.g., by decreasing storage requirements and minimizing spoilage [Ketzenberg and Ferguson, 2008]. Furthermore, sales forecasting enables the analysis of sales drivers such as marketing [Chan and Perry, 2017]. Importantly, the widespread practice of gathering data [Kandel et al., 2012] has made it attractive to many sectors to develop forecasting and inference.

The Bayesian approach to forecasting is gaining popularity due to advances in computational methods making it more viable. Furthermore, Bayesian inference has several appealing benefits in the retail context [Christ, 2011]. Case specific business knowledge can be inserted into models by setting prior distributions, and conclusions in interval estimation are more intuitive with Bayesian probability statements than with the frequentist definition of confidence intervals [Gelman et al., 2013]. For these reasons, we focus on Bayesian methods.

However, in retail sales forecasting it is not obvious how the forecasting model should be constructed. Choosing a model is a complex task that is further discussed by Gelman et al. [2020] and Vehtari and Ojanen [2012]. We focus on a subproblem in model selection called variable selection, which is defined by Piironen et al. [2020] as choosing a minimal subset of explanatory variables such that the predictive performance of the model is good and adding more variables does not significantly improve its performance. Variable selection is an attractive topic in retail sales forecasting because the resulting order of relative importance of explanatory variables can be used in business decision making, i.e., choosing promotions, marketing campaigns or store locations.

We chose the projection predictive method because recent work in several fields has shown interest towards it. Examples are found, for instance,

in ecology [Greenop et al., 2020], public health [Bartonicek et al., 2021], meteorology [Mercer et al., 2020] and psychology [Kelter, 2021].

The projection predictive method provides answers to important business problems in retail sales forecasting by quantifying which explanatory variables are the most important when predicting sales. It also indicates what the smallest subset of explanatory variables is, given a performance level in predicting sales. Answering these questions provides business value by possibly explaining customer behaviour and more importantly, by possibly improving forecasting model performance.

This work was done in a company called Sellforte which develops its own marketing mix model, thereby being interested in accurate forecasting of retail sales. With this perspective in mind, we create a model for explaining the daily total sales of a large grocery retailer and use the projection predictive method to answer the questions of explanatory variable importance and optimal model size that were presented above.

Specifically, we explore popular model and variable selection methods in Section 2, and in Section 3 we present the projection predictive method in detail, along with a retail data set and a Bayesian model for predicting the daily total sales of the data set. Finally, the results are presented in Section 4 and the conclusions are summarized in Section 5.

2. Background

Bayesian modelling is performed by setting up a full probability model, computing and interpreting the conditional probability distributions of model parameters given the observed data, and evaluating the computation process and the implications of the conditional probability distributions [Gelman et al., 2013]. Data scientists often end up repeating these stages in the process of creating a good model [Gelman et al., 2020]. One of the reasons is that there are many ways of defining the full probability model for a problem. Model selection methods are some of the ways with which scientists are able to reduce the number of configurations they have to test before arriving to a satisfactory conclusion.

Bayesian model selection approaches can be categorized by their underlying assumptions of the true data generating mechanism. The categories are called *model views* and the choice of a view determines the model selection process.

- $\mathcal{M}$-closed view assumes that one of the candidate models is the true data generating mechanism.
- $\mathcal{M}$-open view does not assume ability to construct a model which correctly generates the distribution of future data.
- $\mathcal{M}$-completed view assumes the existence of a model M^* which explains the available data with enough accuracy. [Vehtari and Ojanen, 2012]

If the $\mathcal{M}$-closed view can be adopted, the usual practice for model selection would be to calculate the posterior model probabilities

$$p(M|D) \propto p(D|M)p(M) \tag{2.1}$$

for the list of candidate models $\{M_l\}_{l=1}^{L}$ over the model space M and data set D, and select the model with maximum a posteriori (MAP) probability

[Held and Bov, 2013, Kelter, 2021]. Accurate predictions of future data values can also be obtained by Bayesian model averaging (BMA) [Raftery and Zheng, 2003, Piironen and Vehtari, 2017]

$$p(\tilde{y}|D) = \sum_{l=1}^{L} p(\tilde{y}|D, M_l)p(M_l|D) \ . \tag{2.2}$$

However, computing the BMA solution becomes more complicated with a large number of candidate models. Furthermore, the assumptions of the $\mathcal{M}$-closed view can be hard to meet in practice [Kelter, 2021].

In contrast to the $\mathcal{M}$-closed view, it is possible to get away with far less assumptions by adopting the $\mathcal{M}$-open view. Strict assumptions about the true data generating mechanism are then avoided by using samples from the obtained data D as a proxy for the true distribution of future data [Vehtari and Ojanen, 2012]. The idea is used in practice by implementing a cross-validation method or an information criterion estimation, examples of which are given below.

Traditionally, Bayesian K-fold cross-validation (CV) [Geisser and Eddy, 1979] is used, but more modern methods also exist, such as leave-one-out cross-validation with Pareto-smoothed importance sampling (PSIS-LOO-CV), which aims to maximize the amount of cross-validation folds without considerably increasing the computational burden [Vehtari et al., 2015, Vehtari et al., 2017].

Using leave-one-out methods is problematic for estimating predictive performance of time series models because information from future time points $t+1, t+2$ would influence the predictions at time t. Instead, it is possible to use a leave-future-out cross-validation approximation (PSIS-LFO-CV) [Bürkner et al., 2020] or information criterion such as the focused information criterion (FIC) [Pandhare and Ramanathan, 2020].

Finally, the $\mathcal{M}$-completed view combines elements from both the open and closed views. It is also is a compromise in terms of strictness, since it weakens the assumptions of the $\mathcal{M}$-closed view by only requiring an encompassing model that is held as the best available description of future data. Model selection in the $\mathcal{M}$-completed context consists of first creating an encompassing *reference model* in $\mathcal{M}$-closed or $\mathcal{M}$-open context and then selecting one of the candidate models based on how well their predictive performance matches the performance of the reference model [Vehtari and Ojanen, 2012]. Common selection methods, as presented by Piironen [2017], are the reference predictive method and the projection predictive

method, the latter of which is presented in Section 3.3.

This work focuses on the projection predictive method because adopting the $\mathcal{M}$-completed view can sometimes be the most natural option in retail time series forecasting. The assumptions of the $\mathcal{M}$-closed view are often too unrealistic. The $\mathcal{M}$-open view can be a viable choice in many cases but we choose the $\mathcal{M}$-completed view in order to investigate our possibilities in a situation where a gold standard model has already been established. It is often unclear which factors explain sales best, and thus creating a large reference model and then reducing the number of variables could be more effective than other approaches. Furthermore, the projection predictive method was found to perform better compared to other options by Piironen [2017].

3. Methodology

We perform variable selection on a forecasting model of the total sales of an Ecuadorian grocery retailer. In this section we present the sales data set, forecasting model and variable selection method along with discussion about the selected data and model configurations.

The choice of data set is rather significant because the available explanatory variables in the data set determine how meaningful inferences we are able to make out of the results that the projection predictive method gives. In other words, if none of the explanatory variables have significance in explaining the total sales, the relative importance of the variables is not a very meaningful problem to solve. Therefore, we carefully present the properties and possible restrictions of the explanatory variables that were included in the data.

The forecasting model is selected in such a way that it meets the assumptions of our variable selection framework. We have selected the Facebook Prophet model [Taylor and Letham, 2018] which is has several powerful and adaptable features.

3.1 Data set

The Ecuadorian sales data set was chosen mainly due to sufficient amount of data originating from a single city. This precision enables the usage of location dependent data, in this case weather data, because averaging such data from many locations would not represent the underlying phenomenon well. Other reasons for the choice of data set are, for instance, availability of special event dates, category information and explanatory features in the data set.

The data set was donated by Corporación Favorita and it is available on Kaggle as a competition data set [Favorita and Kaggle, 2017]. The data was

restricted using the filters that are specified in Table 3.1. Furthermore, some data points were interpolated by forward filling to preserve data quality in a way that does not inject future information into the past.

Table 3.1. Filters that were applied to the sales data set.

Dimension	Range
Date	2015-05-01 to 2017-03-31
City	Quito
Categories	BEVERAGES, BREAD/BAKERY, CLEANING, DELI, GROCERY I, MEATS, PERSONAL CARE, POULTRY, PRODUCE

The last 60 days of data were held out as a testing data set. Therefore, the range from 2015-05-01 to 2017-01-31 was used for training the forecasting model. The model was trained to fit a sales time series constructed from the total sales of products from the categories in Table 3.1, and it explained the sales using holiday information from the data set and the explanatory variables that are described next. It is worth noting that the earthquake in April 2016 in Ecuador is also listed as a holiday in order to fit the model better despite probable demand outliers immediately after the disaster.

The Corporación Favorita data set contains promotion data and daily oil price data which are used to create explanatory variables. We use the oil price time series as-is and generate the total number of promotions per day as an explanatory variable. In order to obtain more explanatory variables, two other data sets were also merged to the Corporación Favorita data set, but no further feature engineering was performed to edit the explanatory variable time series or to create new ones.

The first additional data set is a monthly food price index that was collected from United Nations Office for the Coordination of Human Affairs (OCHA) [OCHA, 2021]. The data was converted to daily format by interpolating the missing data points between collection days with linear interpolation. The second data set is weather data for the Quito region from the National Oceanic and Atmosphere Administration [NOAA, 2021]. In conclusion, the models in this work are provided with the explanatory variables in Table 3.2 and we attempt to find the optimal subset of them using the methods described in this section.

Table 3.2. Explanatory variables were used for modelling

Category	Data	Type
Store data	Total number of promotions	Daily
Weather data	Precipitation	Daily
	Average temperature	Daily
	Maximum temperature	Daily
	Minimum temperature	Daily
Economic	Oil price	Daily
	Price of potatoes	Monthly
	Price of yellow maize	Monthly

3.2 Facebook Prophet model

Sales data is modeled by fitting a Prophet model that includes a linear trend with changepoints, yearly, monthly and daily seasonalities, holiday components and the explanatory variables. This model is chosen because the different components describe common phenomena in customer behaviour rather well. The model can be expressed with the following equation:

$$y(t) = g(t) + s(t) + h(t) + v(t) + \epsilon_t \ , \tag{3.1}$$

where $g(t)$ represents the linear piecewise trend with non-periodic changepoints, $s(t)$ represents periodic components such as yearly or monthly seasonality, $h(t)$ represents holidays that can occur on irregular intervals over one or more days, $v(t)$ represents the explanatory variables that are time series with daily values and ϵ_t is the error term that represents changes not explained by the model [Taylor and Letham, 2018].

The model presented by Taylor and Letham [2018] does not include the $v(t)$ term in their model specification like Equation (3.1) does. However, within this work it makes sense to add the term since the Prophet library includes capabilities for adding explanatory variables to the model [Facebook, 2021].

Further details of the components of our model are almost identical to the definitions of Taylor and Letham [2018]. They define the trend component of model (3.1) as

$$g(t) = (k + \boldsymbol{a}(t)^T \boldsymbol{\delta})t + (m + \boldsymbol{a}(t)^T \boldsymbol{\gamma}) \ , \tag{3.2}$$

where k is the growth rate, $\boldsymbol{\delta}$ is a vector of rate adjustments, $\boldsymbol{a}(t)$ is a vector determining how the rate has been adjusted at time t. The growth rate is

offset by m, and $\boldsymbol{\gamma}$ is a vector that contains an offset for each time point t. The offsets are calculated from $\boldsymbol{\delta}$ to make the piecewise trend continous.

Seasonalities are modeled by including Fourier series components to the model. For each seasonality period P, the seasonality component is

$$s(t) = \sum_{i=1}^{F} \left(a_i \cos\left(\frac{2\pi i t}{P}\right) + b_i \sin\left(\frac{2\pi i t}{P}\right) \right) , \tag{3.3}$$

where a_i and b_i are the coefficients to be fitted and F is the Fourier order [Taylor and Letham, 2018]. The Fourier order is recommended to be chosen between 1 and 10 according to case specific needs [Facebook, 2021].

Table 3.3 describes the properties of seasonality components that were included in the model in this work.

Table 3.3. Seasonality settings of the model

Seasonality	Period	Fourier order
Monthly	30.5	8
Weekly	7	5

Holidays and continuous explanatory variables are simpler to include in the model since they can be thought of as applying a coefficient to the corresponding holiday occurrence series or explanatory variable time series. The holiday and explanatory variable components are written as

$$h(t) = \sum_{i=1}^{H} c_i \cdot \mathbf{1}(t \in C_i) \tag{3.4}$$

$$v(t) = \sum_{j=1}^{V} d_j \cdot v_j(t) , \tag{3.5}$$

where H and V are the total number of holidays and explanatory variables, respectively. Term $v_j(t)$ is the j:th time series, $\mathbf{1}(t \in C_i)$ indicates if holiday C_i is active on day t and c_i, d_j are the coefficients to be fitted. We use holiday information from the data set but general sets of national holidays are also available.

In conclusion, the model can be represented as

$$\begin{aligned} y &\sim \mathcal{N}((k + A\boldsymbol{\delta})t + (m + A\boldsymbol{\gamma}) + X\boldsymbol{\beta}, 10) \\ k &\sim \mathcal{N}(0, 5) \\ m &\sim \mathcal{N}(0, 5) \\ \delta &\sim \text{Laplace}(0, 0.01) \\ \boldsymbol{\beta} &\sim \mathcal{N}(0, 10) \end{aligned} \tag{3.6}$$

by first separating coefficients from the underlying component functions to vector $\boldsymbol{\beta}$ of length f and giving them a prior distribution. In this case $\boldsymbol{\beta}$ includes the coefficients for seasonalities, holidays and explanatory variables. Then, all individual model components are combined to a $n \times f$ matrix X, where n is the number of days or data points. [Taylor and Letham, 2018]

Many of the explanatory variables are indicators of long term change. Therefore, they compete with the trend component of the model because the trend is also trying to explain long term changes in total sales. To control this and to avoid trend overfitting, the trend changepoint prior was set to a relatively low value of 0.01 and the number of changepoints was restricted to five. On the contrary, other priors were left rather wide because the risk of overfitting is smaller in frequent seasonalities.

3.3 Projection Predictive Framework

The projection predictive framework is defined as a two-stage procedure,

1. Construct the best possible model. This is called the *reference model* and it might be complex for having a large number of variables.

2. If the reference model is too complex, find a simpler model that gives similar predictions compared to the reference model. The simpler model is called a *submodel* and its creation process is referred to as a *projection*. For a given level of complexity (number of variables), the submodel with the smallest predictive discrepancy compared to the reference model is selected. [Piironen et al., 2020]

The predictive discrepancy between a submodel and the reference model is defined as the average Kullback-Leibler (KL) divergence between the predictive distributions of the models. Let the reference model and a submodel be parameterized by $\boldsymbol{\theta}_*$ and $\boldsymbol{\theta}$ respectively. Then, the predictive distributions are $p(\tilde{y}|\boldsymbol{x}, \boldsymbol{\theta}_*)$ and $p(\tilde{y}|\boldsymbol{x}, \boldsymbol{\theta})$, where $\tilde{y}$ represents an unseen data point. In order to minimize discrepancy, the projected submodel parameters are thus defined by

$$\begin{aligned}\boldsymbol{\theta}_\perp &= \arg\min_\theta \frac{1}{n}\sum_{i=1}^{n} \mathbf{KL}(p(\tilde{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}_*)||p(\tilde{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta})) \\ &= \arg\min_\theta \frac{1}{n}\sum_{i=1}^{n} p(\tilde{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}_*) \cdot \log\left(\frac{p(\tilde{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}_*)}{p(\tilde{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta})}\right) \ , \end{aligned} \tag{3.7}$$

where the equivalence follows from the definition of KL divergence. As long as the observation model for y_i belongs to the exponential family, projection (3.7) is equivalent to finding the maximum expected likelihood parameters for $\boldsymbol{\theta}$ with the original observations y_i replaced by their expected values $\mathrm{E}(y_i|\boldsymbol{x_i}, \boldsymbol{\theta_*})$ over the distribution $p(\boldsymbol{\theta}_*|\mathcal{D})$ as predicted by the reference model. [Piironen et al., 2020]

As described in Section 3.2 where the selected model was defined, the observation model for y_i is a normal distribution. Therefore, the model belongs to the exponential family and the original observations can be replaced by the predictions of the reference model.

To obtain the posterior distribution of a submodel, the projection (3.7) is performed several times. By projecting S draws $\{\boldsymbol{\theta}_*^s\}_{s=1}^S$ from the posterior $p(\boldsymbol{\theta_*}|\mathcal{D})$ we obtain corresponding draws $\{\boldsymbol{\theta}^s\}_{s=1}^S$ in the projection space. These projected draws form the posterior distribution of the submodel. [Piironen et al., 2020]

Possible submodels are explored by forward search to decrease the amount of projections that have to be computed. Even faster L1-like heuristic methods exist but they are not necessary in this case. The forward search means first projecting to an intercept model (no explanatory variables) and then sequentially adding the variable which decreases discrepancy to the reference model predictions the most. This is a greedy approach that avoids checking many clearly non-optimal combinations whilst still finding good solutions in practice. [Piironen et al., 2020]

Given the variable combinations and thus submodels for each size, the final model selection is performed by choosing the submodel that has predictive performance close enough to the predictive performance of the reference model. It is common to consider 95% match in performance to be sufficient. Predictive performance is measured by mean absolute percentage error (MAPE) which is defined as

$$\mathrm{MAPE} = \sum_{i=1}^{n} \frac{1}{n} \left| \frac{y - \tilde{y}}{y} \right| \tag{3.8}$$

and expected log predictive density (ELPD) which is defined as

$$\mathrm{ELPD} = \sum_{i=1}^{n} \int p(\tilde{y}_i) \log(p(\tilde{y}_i|y)) d\tilde{y}_i \ . \tag{3.9}$$

4. Results

In this section we present the results of fitting the reference model to the training data set and then projecting the information to submodels of different sizes while using forward search to select the best combination of explanatory variables for each size. The predictive performance of the submodels is then evaluated against the performance of the reference model on the training and test sets using mean absolute percentage error (MAPE) and expected log predictive density (ELPD), which are defined in Equations (3.8) and (3.9), respectively. Finally, model selection is performed by selecting the submodel that has as few explanatory variables as possible while matching reference model performance with 95% accuracy.

The reference model was fit to the training data set using Markov chain Monte Carlo (MCMC) sampling with 400 samples. As explained in Section 3, the training set consists of two years of sales data and the test set is 60 days of sales data beginning from the day after the training data ends. The amount of sales and items on promotion were log-transformed in order to help model fitting. Predictions of the obtained model against the training data points are plotted in Figure 4.1. Figures of fitted values for individual model components can be found in Appendix A.

4.1 Order of importance for the explanatory variables

After the reference model was fitted, submodels were collected using forward search. Submodels are hereon referred by their size, which means the number of explanatory variables in the submodel. Submodel of size n has the first n variables added by the forward search.

From Figure 4.2, it can be observed that when increasing the number of explanatory variables in a submodel, the posterior predictive distribution of the submodel approaches the posterior predictive distribution of the

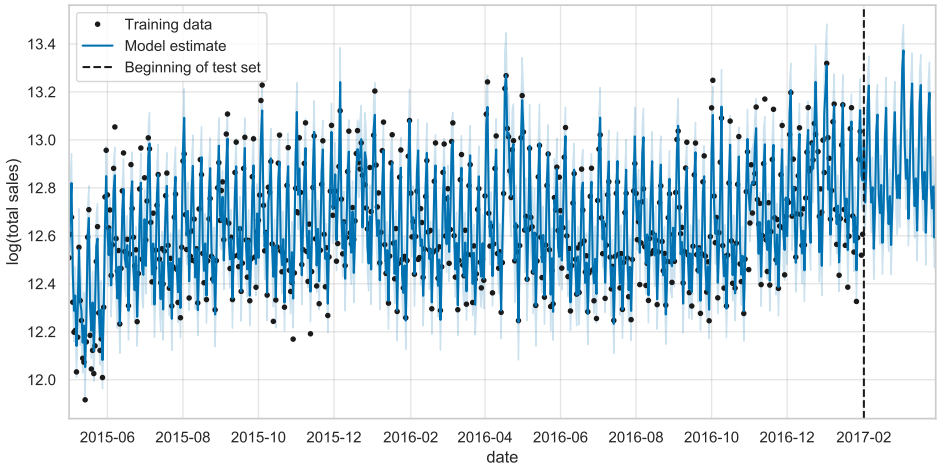

Figure 4.1. Reference model is fitted to the data. Model predictions seem reasonable visually but on the contrary, some data points are clearly not fitted well.

reference model. This is indicated by the discrepancy (KL divergence) approaching zero, which indicates that the search process is satisfactory. Therefore, we can conclude that according to the order in which the variables were added by the forward search process, the order of importance for the explanatory variables is:

1. Price of potatoes
2. Average temperature
3. Minimum temperature
4. Oil price
5. Total number of promotions
6. Precipitation
7. Maximum temperature
8. Yellow maize price.

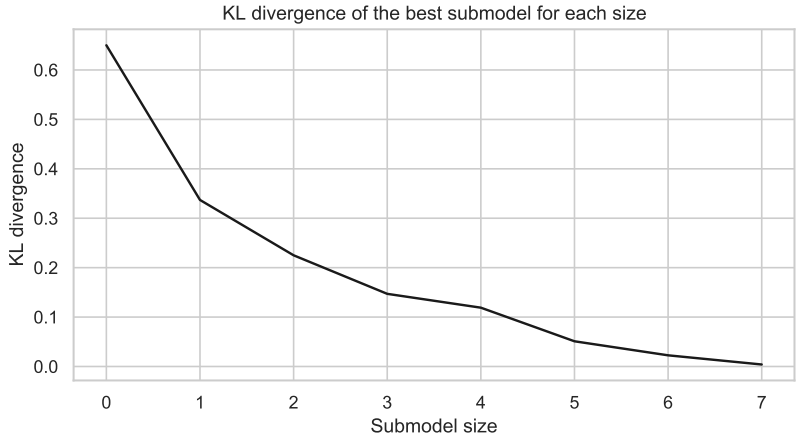

Figure 4.2. KL divergence of the predictive distributions of the best submodel and the reference model decreases as the size of the submodel is increased.

Possible reasons for this particular order of variables can be found by

analyzing the time series of all explanatory variables in Figure 4.3. For example, the low ranking of maximum temperature is likely caused by the high correlation between maximum and average temperature time series. Adding the maximum temperature time series does not help in fitting the model if it is nearly equivalent to adjusting the coefficient of average temperature. A similar reasoning could also explain the low importance of yellow maize prices compared to high ranking of the price of potatoes. Visually, they seem to have very high inverse correlation which has similar problems as high positive correlation.

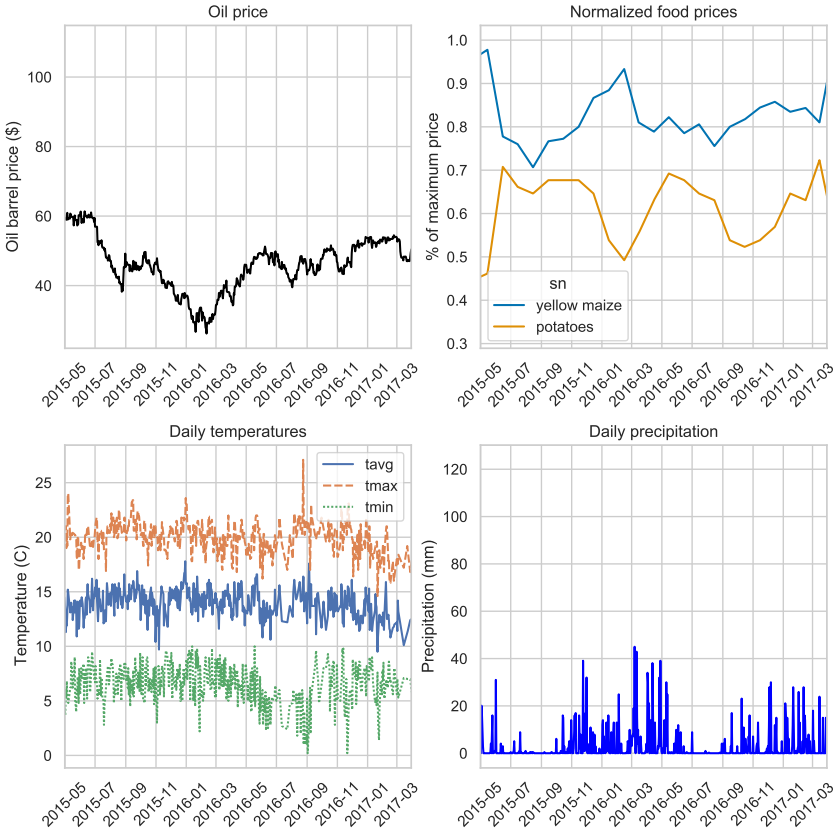

Figure 4.3. Time series plots of the explanatory variables help explain their relative importance.

The relative order of importance could also be affected by model structure. The trend component of our model tries to explain all non-seasonal long-term changes and seasonal effects with season length longer than one month. Therefore, by setting the maximum number of trend changepoints to five, we intentionally avoid overfitting and could end up selecting more explanatory variables that include some long-term changes that the trend is not flexible enough to explain. It can be argued that the four most

important variables all represent some type of long-term seasonal or non-seasonal change: temperatures change on an annual cycle while prices of oil and potatoes can have certain non-annual changes. Figure 4.4 displays how the trend component of the smaller models, especially size 0, deviates from the reference model trend component.

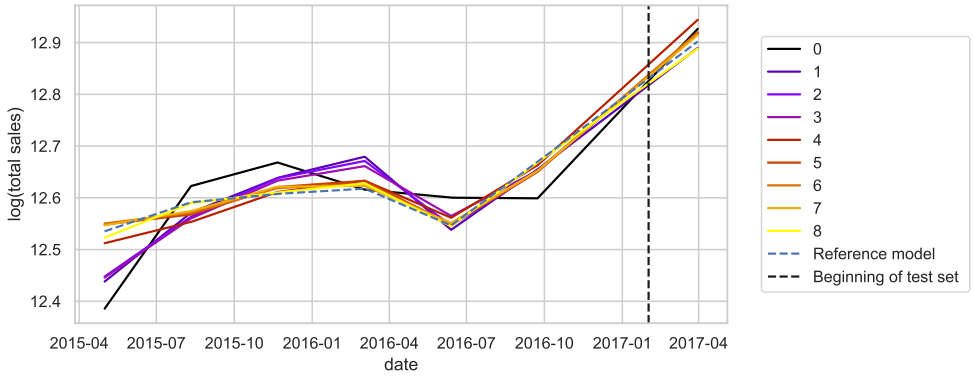

Figure 4.4. The trend component of the model explains non-seasonal long-term changes and seasonal changes with a period longer than 30 days. Trend components of the smaller models deviate from the reference model trend, indicating that the trend is compensating for the lack of explanatory variables.

4.2 Selecting the optimal submodel

Model selection is performed by plotting the predictive accuracy of the reference model and each of the submodels that were found by the forward search process. We then select the one that has as few variables as possible but matches reference model performance well. Performance is measured by using MAPE and ELPD, which are defined in Equations (3.9) and (3.8), respectively.

Figure 4.5 displays the selected performance statistics against the observed data over the training data set. Reference model performance, is matched with 95% tolerance in MAPE and ELPD already at size one. The performance is slightly improved in larger submodels but the increases are negligible in comparison to the uncertainty of the metric.

Using the held out test data set, it is possible to also evaluate the predictive performance on actual unseen observations. The predictive performance over the 60-day test data set is shown in Figure 4.6. Unlike in the training data, the worst performing model seems to be the size 4 model while reference model performance is matched or exceeded already with the size 0 submodel. This is counterintuitive under the assumption that the reference model is the best available description of the data.

Since we do not test alternative assumptions within the scope of this

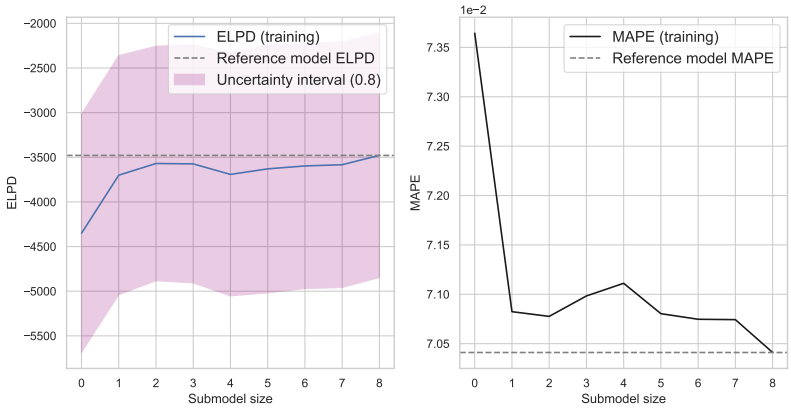

Figure 4.5. Performance of the submodels matches reference model performance already with one explanatory variable when measuring over the training data.

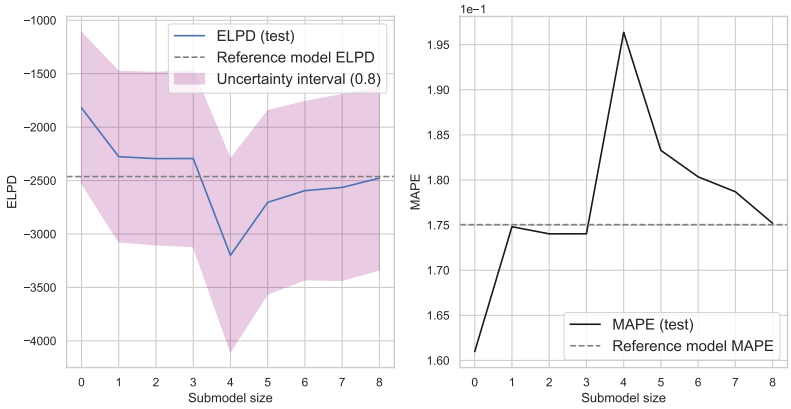

Figure 4.6. Performance statistics over the test data set are inconclusive of the correct submodel size since they violate our assumptions and are under significant uncertainty.

work, we settle for concluding we are unable to assess the predictive performance for unseen data due to uncertainty in the test data and difficulties in the predictive performance of the reference model. As seen in the Figure 4.7, both the reference model and the size 1 submodel have difficulties in predicting some data points, especially in the test set.

From Figure 4.7, we observe that the selected submodel mean fits rather well to the original data, although some systematic error can be observed in the test data portion. We can also see that the predictive distribution of the submodel is slightly narrower than the distribution of the reference model. This could be a side effect of the estimation process that was used in projecting the reference model draws.

In conclusion, the results indicate that the predictive performance of the reference model can be approximately matched with a submodel that has only potato price as an explanatory variable. However, the reference model performance has room for improvement, since we do not see clear

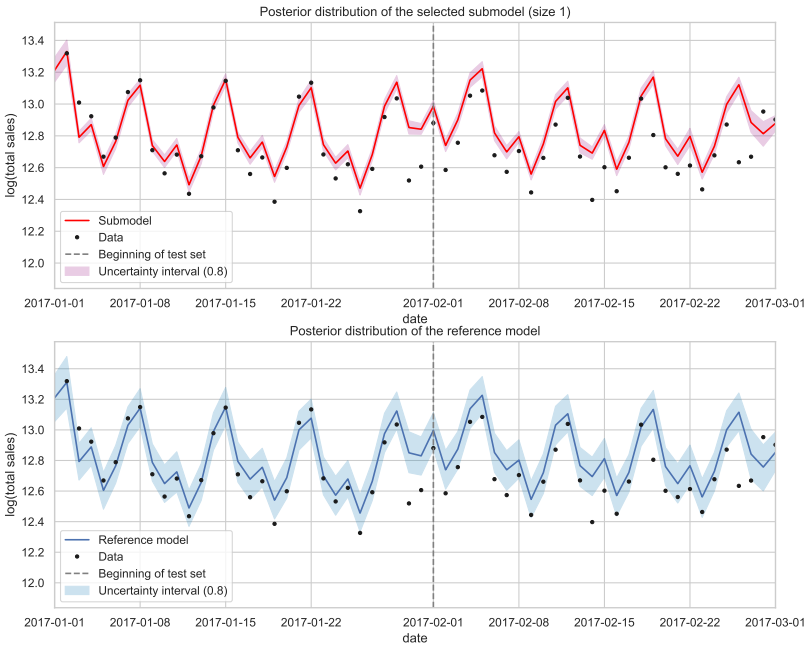

Figure 4.7. Plot of the posterior predictive distributions of the selected submodel and the reference model.

improvement of performance over the test set or narrow uncertainty for performance over the training data set.

5. Conclusions

In this thesis, we modeled the daily total sales of a grocery retailer using a Bayesian forecasting model. We applied the projection predictive method to reduce the number of explanatory variables in the model without significantly decreasing model performance. In the light of our experiment, the projection predictive method was an excellent method for variable selection because it provided us with quantitative information about the relative importance of explanatory variables. Furthermore, it indicated that in our experiment, a model with only the price of potatoes as an explanatory variable achieves similar predictive performance as a model that has also the other seven variables included.

Unfortunately, the experiment design posed some challenges. Relative to the ELPD uncertainty, there were not many variables that significantly increased the predictive performance of the model. The projection predictive method would likely give stronger results if there were more explanatory variables under consideration, and more importantly, more variables that have significant predictive power.

Further studies could focus on improving the experiment by creating more explanatory variables, for example, by creating boolean variables that indicate some conditions such as heavy rain or a hot day. More product category level information could also be included, such as the number of promotions for each category. Using a vast range of detailed promotion time series could give impactful insights especially from the retailer's perspective.

The $\mathcal{M}$-complete assumption could also be questioned by experimenting with variable and model selection in the $\mathcal{M}$-open context which could also be applicable for sales forecasting. That would, however, require significantly more resources since structurally different models must also be tested in addition to varying the number of explanatory variables.

Regardless of the limitations of this experiment, the evidence suggests that the projection predictive method is an applicable variable selection method to modelling retail sales with the Facebook Prophet modelling framework.

Bibliography

A. Bartonicek, S. R. Wickham, N. Pat, and T. S. Conner. The value of bayesian predictive projection for variable selection: an example of selecting lifestyle predictors of young adult well-being. *BMC public health*, 21(1):695, Apr 09 2021.

Paul-Christian Bürkner, Jonah Gabry, and Aki Vehtari. Approximate leave-future-out cross-validation for bayesian time series models. *Journal of statistical computation and simulation*, 90(14):2499–2523, 2020.

David Chan and Mike Perry. Challenges and opportunities in media mix modeling. 2017.

Steffen Christ. *Operationalizing Dynamic Pricing Models: Bayesian Demand Forecasting and Customer Choice Modeling for Low Cost Carriers*. Gabler Verlag, Wiesbaden, 1. aufl. edition, 2011.

Facebook. Prophet: Forecasting at scale. `https://github.com/facebook/prophet`, 2021. Version 1.0.

Corporación Favorita and Kaggle. Corporación favorita grocery sales dataset. `https://www.kaggle.com/c/favorita-grocery-sales-forecasting/data`, 2017. Accessed: 2021-07-26.

Seymour Geisser and William F. Eddy. A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365):153–160, 1979.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, 3 edition, 2013.

Andrew Gelman, Aki Vehtari, Daniel Simpson, Charles C Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. Bayesian workflow. *arXiv preprint arXiv:2011.01808*, 2020.

Arran Greenop, Samantha M. Cook, Andrew Wilby, Richard F. Pywell, and Ben A. Woodcock. Invertebrate community structure predicts natural pest control resilience to insecticide exposure. *Journal of Applied Ecology*, 57(12):2441–2453, 2020.

Leonhard Held and Daniel Sabans Bov. *Applied Statistical Inference: Likelihood and Bayes*. Springer Publishing Company, Incorporated, 2013.

Sean Kandel, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer. Enterprise data analysis and visualization: An interview study. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2917–2926, 2012.

Riko Kelter. Bayesian model selection in the m-open setting — approximate posterior inference and subsampling for efficient large-scale leave-one-out cross-validation via the difference estimator. *Journal of Mathematical Psychology*, 100:102474, 2021.

Michael Ketzenberg and Mark E Ferguson. Managing slow-moving perishables in the grocery industry. *Production and Operations Management*, 17(5):513–521, 2008.

Jason J. Mercer, David T. Liefert, and David G. Williams. Atmospheric vapour and precipitation are not in isotopic equilibrium in a continental mountain environment. *Hydrological Processes*, 34(14):3078–3101, 2020.

NOAA. Climate data online. `https://www.ncdc.noaa.gov/cdo-web/`, 2021. Accessed: 2021-07-26.

OCHA. Ecuador - food prices. `https://data.humdata.org/dataset/wfp-food-prices-for-ecuador`, 2021. Accessed: 2021-07-26.

S. C. Pandhare and T. V. Ramanathan. The focussed information criterion for generalised linear regression models for time series. *Australian & New Zealand Journal of Statistics*, 62(4):485–507, 2020.

Juho Piironen and Aki Vehtari. Comparison of bayesian predictive methods for model selection. *Statistics and computing*, 27(3):711–735, 2017.

Juho Piironen, Markus Paasiniemi, and Aki Vehtari. Projective inference in high-dimensional problems: Prediction and feature selection. *Electronic Journal of Statistics*, 14(1):2155–2197, 2020.

Adrian E Raftery and Yingye Zheng. Discussion: Performance of bayesian model averaging. *Journal of the American Statistical Association*, 98(464):931–938, 2003.

Sean J Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018.

Aki Vehtari and Janne Ojanen. A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6(none):142 – 228, 2012.

Aki Vehtari, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry. Pareto Smoothed Importance Sampling. *arXiv e-prints*, art. arXiv:1507.02646, July 2015.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 27(5):1413–1432, 2017.

A. Appendix

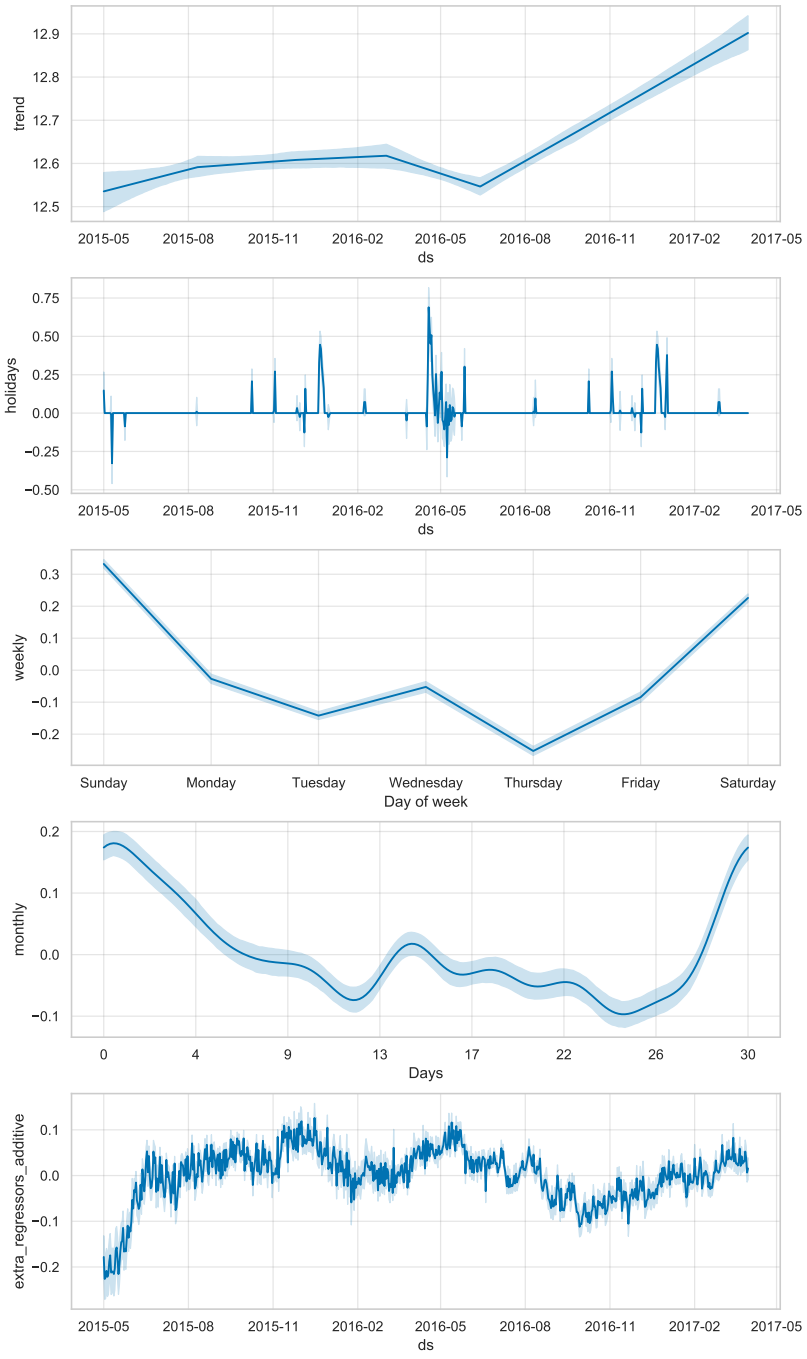

Figure 1.1. Components of the Prophet reference model.