Demand Forecasting of individual Probability Density Functions with Machine Learning

Felix Wick [*1], Ulrich Kerzel [†2], Trapti Singhal [‡1], and Martin Hahn [§1]

[1]Blue Yonder GmbH (Karlsruhe, Germany)
[2]IUBH Internationale Hochschule (Erfurt, Germany)

Demand forecasting is a central component for many aspects of supply chain operations as it provides crucial input for subsequent decision making like ordering processes. While machine learning methods can significantly improve prediction accuracy over traditional time series forecasting, the calculated predictions are often mere point estimations for the conditional mean of the underlying probability distribution, and the most powerful approaches, like deep learning, are usually opaque in terms of how its individual predictions can be interpreted. Using the novel supervised machine learning method "Cyclic Boosting", complete individual probability density functions in form of negative binomial distributions can be predicted instead of simple point estimates. While metrics evaluating simple point estimates are widely used, methods for assessing the accuracy of predicted distributions are rare and this work proposes new techniques for both qualitative and quantitative evaluation methods. Additionally, each single prediction obtained with this framework is explainable, which is a major benefit in practice as individual forecasts can be understood by the practitioner and "black-box" models can be avoided.

Keywords: **explainable machine learning, demand forecasting, probability distribution**

1. Introduction

Demand forecasting is one of the main challenges for retailers and at the core of business operations. Due to its stochastic nature, demand is difficult to forecast as it depends on many influencing factors and the realized demand can be interpreted as a random variable that is described by an appropriate probability density function (PDF). In order to make operational decisions, an optimal point estimator has to be defined, that can be used to derive ordering decisions used in the replenishment process of the retailer. Demand estimation is further complicated by the fact that retailers typically only observe realized sales rather than the actual demand, and in case the demand exceeds the current stock level the data become censored. The ordering decision process is complicated by a range of factors: Even in the case of perfect demand forecasts, the decision maker has to consider lot-sizes defined by the wholesaler or manufacturer as well as to balance conflicting metrics to reach an optimal decision: Ordering too few items may result in stock-out situations leading to unrealized demand and unsatisfied customers. Ordering too many items results in excess inventory which increases transport and storage costs and, in the case of perishable goods, excessive waste, as spoilt items need to be disposed of at additional cost and potentially even environmental impact. This situation is particularly noticeable in the so called "ultra-fresh" category, which includes items such as bakery products, ready-meals, fresh diary products, or certain meat products such as ground meat. These items typically have a shelf-life ranging from less than a business day to a few business days at most, with a continuous spectrum in between, depending on the exact item. In many situations, additional constraints have to be considered to reach an optimal ordering decision: Delivery cycles of items may vary depending on the type of item and the wholesaler or manufacturer from which they are procured. Retailers also operate at a given service level to guarantee that a certain level of demand can be fulfilled. The exact service level typically depends on the overall business policy of the retailer and may also depend on individual products, ranging from "never-out-of-stock" items to a service level exceeding e.g. 90%.

The availability of Big Data allows capturing, storing and processing a vast amount of data associated with demand, such as historic sales records, information about promotional events or advertisements, pricing information, local weather at retail locations, seasonal information as well as a wide range of further variables. Modern machine learning algorithms can then be used to predict the per-item demand distribution, corrected for censored data, from which an optimal point estimator can be derived to be used in the subsequent ordering decision. It is important to note that demand as a random variable is not identically and independently distributed (i.i.d.). While the probability distribution describing the demand can be attributed to a given family or parameterization, the exact parameters vary: Seasonal effects, finite life cycles of products and the introduction of new products influence

[*]felix.wick@blueyonder.com
[†]u.kerzel@iubh-fernstudium.de
[‡]trapti.singhal@blueyonder.com
[§]martin.hahn@blueyonder.com

the demand distribution, as well as the local weather at the retail location or the retail location itself in terms of size, assortment range, customer diversity and other factors. The retailers themselves also actively influence demand by using advertisements to highlight products, offering rebates or discounts for specific products as well as pursuing an active pricing strategy. This means that while we can generally assume that demand follows a specific type of probability distribution, its parameters are unique to the instance for which an estimate is required. For example, the probability distribution governing the demand of a particular item is specific to the item, date and retail location for which the forecast is made and depends on a wide range of further influencing factors.

The remainder of the paper is organized as follows: We first review the relevant literature and existing work in sec. 2. We then describe our method to predict individual negative binomial PDFs by means of a parametric approach including two distinct machine learning models for mean of variance in sec. 3. After that we describe methods for the qualitative and quantitative evaluation of PDF predictions in sec. 4. And finally, we present a demand forecasting example to show an application of our methods in sec. 5.

2. Literature Review

Inventory management offers a rich theory and the extensive body of research can be broadly grouped into two categories, where the inventory control problem is either based on some knowledge of the underlying demand distribution or an integrated approach that seeks to map directly from the available data (historic sales records and further variables) to the ordering decision. This approach is often referred to as "data-driven newsvendor" and discussed e.g. in [1–4]. It aims to avoid estimating the underlying probability distribution for demand and use the available data to derive the operational decisions (the order quantity) directly. An overview of a range approaches can also be found in [5].

Although this approach seems preferable at first glance, since it avoids determining the full demand distribution and results directly in the desired operational decision (the order quantity), it faces several drawbacks. First, the full probability distribution for the demand of a specific item at a given sales location and business day includes all available information including the uncertainty of the modelled demand. This can be used to simulate the performance on a per-item level and e.g. optimize the impact on business strategy decisions on conflicting metrics such as stock out- and waste-rate. By forecasting the full demand distribution as opposed to point estimators such as the expected demand or the median of the distribution, the forecast quality can be evaluated for quantiles of the distribution, including the often extensive tails of the distribution. Additionally, a purely data-driven approach going from the observed data directly to the operational decision (such as the order quantity) does not allow to analyze the data-generating process, i.e. the mechanism behind the stochastic behavior of the customer demand. However, modelling demand directly is vital if a causal analysis is planned at a later stage or independently, for example to study the effect promotions, advertisements, price changes or other influencing factors in either Pearl's do-calculus [6] or Rubin's potential outcomes framework [7]. Using an operational quantity such as the order quantity will in most cases act as an insufficient proxy of the quantity of interest (customer demand) and likely lead to unnecessary causal pathways that may not be able to be fully controlled for. From a practical perspective, separating the demand forecast from the operational decisions (i.e. calculating the order quantities for the next delivery cycle) also allows to evaluate longer-term planning and reduces the complexity as it avoids coupling the complex delivery schedules of multiple wholesalers and manufacturers with the forecast of customer demand. This also allows to share long-term demand predictions with other business units or external vendors and wholesalers to ease their planning for the production and supply-chain processes upstream of the retailer. From the perspective of industrial practice, modelling the demand separately from deriving the subsequent orders has the additional benefit that multiple retail chains can benefit from any improvement in the model description even if the concrete retailers are unrelated to each other. For example, if a particular effect was identified at some retailer A and included in the machine learning model, the improvement can be rolled out immediately or on request to all other retailers using this system on a planetary scale without having to adapt the underlying machine learning model for each retailer individually.

In contrast to the direct mapping the observed data to ordering decisions, more traditional inventory control systems rely on the knowledge of the demand distribution in one form or another, see e.g. [8] for an overview. In (s, S) type inventory control systems [9], inventory levels are monitored at regular intervals and orders are dispatched once the inventory level reaches a minimal value s. In case of linear holding and shortage costs, such policies are optimal [10], although perishable goods pose more challenges, see e.g. [11–13]. Additionally, service level constraints can be included in these kind of inventory control systems [14]. Perishable goods are well described by the "newsvendor-problem" [15], where in the simplest case all stock perishes at the end of the selling period (e.g. a business day). For a detailed review of the newsvendor problem see e.g. [16]. Assuming linear underage and overage costs $b, h > 0$, the optimal quantile $q_{\text{opt}} = b/(b+h)$ of a known demand distribution $f(D)$ can be calculated exactly.

The main objective in any of these approaches is to determine the underlying demand distribution. The simplest approach is to just use the observed sales events and forecast these as a time series (see e.g. [17]) or via sample average approximation (SAA), see e.g. [18] for an overview. However, these approaches do not make use of any data apart from the sales record themselves, although we know that many variables such as price or advertisements influence, and therefore are highly correlated with, the demand. Saghafian and Tomlin [19] propose to include partial information about the distribution in the derivation of the operational decision, i.e. the calculation of the optimal order quantity.

However, in order to be able to fully optimize the opera-

tional decision, it is critical that one indeed reconstructs a full demand distribution. This also implies that a simple point-estimator, as provided by the most common statistical techniques and machine-learning approaches, will not suffice. Additionally, we need to consider that demand is not i.i.d., but depends on external factors such as season, weather, product life-cycle, and is also actively changed by the retailer by setting a specific price, offering rebates or running advertisements. Additionally, the demand implicitly depends on the location of the retail outlet as well as the specifics of that location, such as product assortment influencing the choice of possible replacement articles and many more. These complications are the main reason we cannot treat the replenishment process as n independent newsvendor-type problems.

Instead, we need to determine the full demand distribution from data, conditional on the relevant variables such as date, location, and item, taking all auxiliary data such as article characteristics, pricing, advertisements, retail location details, etc. into account. This can be done in several ways: Quantile regression [20, 21] can be implemented in various frameworks and used to estimate a range of quantiles for each predicted distribution from which an empirical probability distribution can be interpolated. Using a dedicated neural network [22], either the full probability distribution or a defined range of quantiles can be calculated directly from the data for each individual prediction without assuming an underlying model. Alternatively, one can assume a given demand model and fit the model parameters instead of reconstructing the complete distribution [23, 24]. This approach is computationally favourable, as fewer parameters need to be estimated compared to the case of the full distribution. Empirically, one can determine the best fitting distribution from data [25]. However, given the stochastic nature of the demand, such an empirically determined distribution is not expected to be stable and prone to sudden changes. Instead, the choice of the demand distribution should be motivated by theoretic considerations. The discrete demand is typically modelled as a negative binomial distribution (NBD), also known as Gamma-Poisson distribution [26–30]. This distribution arises if the Poisson parameter μ is a random variable itself that follows a Gamma distribution. The NBD has two parameters, μ and $\sigma^2 > \mu$, and is overdispersed compared to the Poisson distribution for which $\mu = \sigma^2$. Hence, for each ordering decision, the model parameters μ and σ need to be determined for each item at the required granularity, typically for each sales location and ordering time, depending on all auxiliary data describing article details, retail location, and influencing factors such as pricing and advertisement information.

Summary of Contributions

This work demonstrates how the explainable machine learning algorithm Cyclic Boosting [31] can be used to model the demand distribution at the granularity needed by the retailer. Typically this means that the full demand distribution has to be estimated per SKU for each sales location and opening day, conditional on a wide range of variables such as weather, prices, promotions, etc. In contrast to using a "black-box" machine learning model, using Cyclic Boosting allows to interpret how each individual prediction was made and what the most important variables are per individual prediction.

Additionally, we show how the cumulative distribution function (CDF) can be used to accurately asses the forecast quality of the full predicted demand distribution, including the tails of the distribution. This allows to verify that the predicted demand distribution accurately reflects the observed data and can hence be used both to derive operational decisions such as order quantities as well as strategic business decisions by the retailer or gain further insights into customer behavior using for example causal modelling.

3. Negative Binomial PDF Estimation

To predict an individual PDF using a parametric approach, one has to rely on a model assumption about the underlying distribution of the random variable to be predicted. As discussed earlier, the negative binomial probability distribution (NBD) is well routed in theoretical arguments to model customer demand. Its parameters can be modelled by two independent models, one to estimate the mean and the other for the variance. At least in principle, any method can be used. However, in the case of demand forecasting, each prediction is highly specific to the circumstances in which it is used (such as SKU, opening day, and store location) and may depend on a multitude of describing variables or features such as sales location, weather, price, and so forth. It should be noted that these parameters are not independent between products. For example, a promotion applied to one product can, at least in principle, affect the sales of related products within the assortment. This implies that the demand forecasts cannot be treated as individual newsvendor-type predictions but need to be modelled holistically. Machine learning algorithms are ideally suited for this task and in the following we will use the Cyclic Boosting algorithm to benefit in particular from explainable decisions rather than black-box approaches. Furthermore, the regularization approach used during training of the Cyclic Boosting algorithm allows a dedicated treatment of the underlying NBD model, which is another major benefit compared to a standard "off-the-shelf" machine learning algorithm. This means we use two subsequent Cyclic Boosting models in order to estimate the parameters of each individual PDF that we need to forecast, where the first model is used to estimate the mean and the second to estimate the variance. The features may or may not differ between the mean and variance estimation models. The assigned mean and variance predictions can then be used to generate individual PDFs using the parameterization of the NBD for each sample.

In the following, after a brief recap of the fundamental ideas of Cyclic Boosting, we describe a method to predict mean and variance for individual NBD models using Cyclic Boosting.

3.1. Cyclic Boosting Algorithm: Mean estimation

Cyclic Boosting [31] is a type of generalized additive models using a cyclic coordinate descent optimization and

featuring a boosting-like update of parameters. Major benefits of Cyclic Boosting are its accuracy, performance even at large scale and providing fully explainable predictions which are of vital importance in practical applications, in particular for multi-national retail vendors.

The main idea of this algorithm is the following: First, each feature, denoted by index j, is discretized appropriately into k bins to reflect the specific behaviour of the feature. The global mean μ is determined from all target values y of the target variable $Y \in [0, \infty)$ observed in the data. Single data records, for example the sales of a specific SKU along with all relevant features, are indexed by i. The individual predictions $\hat{y}_i$ can then be calculated as:

$$\hat{y}_i = \mu \cdot \prod_{j=1}^{p} f_j^k \quad \text{with } k = \{x_{j,i} \in b_j^k\} \tag{1}$$

The factors f_j^k are the model parameters that are determined iteratively from the features, and are determined iteratively until the algorithm converges. During training, regularization techniques are applied to avoid overfitting and improve the generalization ability of the algorithm. The deviation of each factor from $f_j^k = 1$ can then be used to explain how a specific feature contributes to each individual prediction.

In detail, the following meta-algorithm describes how the model parameters f_j^k are obtained from the training data:

1. Calculate the global average μ from all observed y across all bins k and features j.

2. Initialize the factors $f_j^k \leftarrow 1$

3. Cyclically iterate through features $j = 1, ..., p$ and calculate in turn for each bin k the partial factors g and corresponding aggregated factors f, where indices t (current iteration) and τ (current or preceding iteration) refer to iterations of full feature cycles as the training of the algorithm progresses:

$$g_{j,t}^k = \frac{\sum\limits_{x_{j,i} \in b_j^k} y_i}{\sum\limits_{x_{j,i} \in b_j^k} \hat{y}_{i,\tau}} \quad \text{where} \quad f_{j,t}^k = \prod_{s=1}^{t} g_{j,s}^k \tag{2}$$

   Here, g is a factor that is multiplied to f_{t-1} in each iteration. The current prediction, $\hat{y}_\tau$, is calculated according to eqn. 1 with the current values of the aggregated factors f:

$$\hat{y}_{i,\tau} = \mu \cdot \prod_{j=1}^{p} f_{j,\tau}^k \tag{3}$$

   To be precise, the determination of $g_{j,t}^k$ for a specific feature j uses $f_{j,t-1}^k$ in the calculation of $\hat{y}$. For the factors of all other features, the newest available values are used, i.e., depending on the sequence of features in the algorithm, either from the current ($\tau = t$) or the preceding iteration ($\tau = t - 1$).

4. Quit when stopping criteria, e.g. the maximum number of iterations or no further improvement of an error metric such as the mean absolute deviation (MAD) or mean squared error (MSE), are met at the end of a full feature cycle.

3.2. Cyclic Boosting Algorithm: Width estimation

In the previous section, the general Cyclic Boosting algorithm was used to estimate the mean of the NBD model. In order to predict the variance of the NBD model (associated with the mean predicted before), we modify the algorithm like described in the following.

When looking at the demand of individual SKUs, the target variable y has the values $y = 0, 1, 2, ...$ and the NBD model can be parameterized as in [32]:

$$\mathrm{NB}(y; \mu, r) = \frac{\Gamma(r + y)}{y! \cdot \Gamma(r)} \cdot \left(\frac{r}{r + \mu}\right)^r \cdot \left(\frac{\mu}{r + \mu}\right)^y, \tag{4}$$

where μ is the mean of the distribution and r a dispersion parameter.

By bounding the inverse of the dispersion parameter $1/r$ to the interval $[0, 1]$ (corresponding to bounding r to the interval $[1, \infty]$), the variance σ^2 can be calculated from μ and r via:

$$\sigma^2 = \mu + \frac{\mu^2}{r} \tag{5}$$

The estimate of the dispersion parameter $\hat{r}$ can then be calculated by minimizing the loss function defined in Eqn. (6), which is expressed as negative log-likelihood function of a negative binomial distribution. Using Cyclic Boosting, the minimization over all input samples i is performed with respect to the Cyclic Boosting parameters f_j^k, constituting the model of $\hat{r}_i$, according to Eqn. (7), where the estimates for the mean $\hat{\mu}_i$ are fixed to the values obtained in the previous step (described in sec. 3.1).

$$L(r) = -\mathcal{L}(r) = -\ln \sum_i \mathrm{NB}(y_i; \hat{\mu}_i, \hat{r}_i) \tag{6}$$

$$\hat{r}_i = 1 + \prod_{j=1}^{p} f_j^k \quad \text{with } k = \{x_{j,i} \in b_j^k\} \tag{7}$$

In other words, the values $\hat{r}_i$ are estimated via learning the Cyclic Boosting model parameters f_j^k for each feature j and bin k from data. For any concrete observation i, the index k of the bin is determined by the value of the feature $x_{j,i}$ and the subsequent look-up into which bin this observation falls. Like in sec. 3.1, the model parameters f_j^k correspond to factors with values in $[0, \infty]$ and again values deviating from $f_j^k = 1$ can be used to explain the relative importance of a specific feature contributing to individual predictions. Note that the structure of Eqn. (7) can be interpreted as inverse of a logit link function in the same way as explained in [31] when Cyclic Boosting is used for classification tasks.

The Cyclic Boosting algorithm is trained iteratively using cyclic coordinate descent, processing one feature with all its bins at a time until convergence is reached. Unlike in the basic multiplicative regression mode of Cyclic Boosting described in sec. 3.1, the minimization of the loss function in Eqn. (6) cannot be solved analytically and has to be done numerically, for example using a random

search. All other advantages of Cyclic Boosting, like for example individual explainability of predictions, remain valid for its negative binomial width mode.

Finally, the variance $\hat{\sigma}_i^2$ can be estimated from the dispersion parameter $\hat{r}_i$ using Eqn. (5). Using the individual predicted mean $\hat{\mu}_i$ from the first step, the model is fully specified for each individual prediction i.

4. Evaluation of PDF Predictions

Many statistical and most machine learning methods do not provide a full probability density distribution as result. Instead, these methods typically predict a single numerical quantity that is then compared to the observed concrete realization of the random variable using metrics such as the mean squared error (MSE), mean absolute deviation (MAD) or others. In the setting of a retailer, the observed quantity is the sales of individual products (denoted by y) and most machine learning approaches would then predict a single number $\hat{y}$ as a direct estimate of the sales. However, reducing the prediction to a single number does not allow to account for the uncertainty of the prediction or the dynamics of the system. Instead, it is imperative to predict the full PDF for each prediction to be able to optimize the subsequent operational decision. Unfortunately, most statistical or machine learning methods that predict full individual probability functions lack quantitative or at least qualitative evaluation methods to assess whether the full distribution has been forecast correctly, in particular in the tails of the distribution.

For an estimation of the determining parameters of an assumed functional form for the PDF, assessing the correctness of the PDF model output refers to the evaluation of the accuracy of the prediction of the different determining parameters. In the case of the negative binomial distribution used in this work, we have to verify that mean and variance are determined accurately, as well as checking that the choice of the underlying model can describe the observed data.

In the following, we will show how different visualizations of the observed cumulative distribution function (CDF) values can be used to evaluate the quality of the predicted PDFs. Although we limit the following discussion to the negative binomial model, the method can be applied generally to any representation of a probability density distribution, even if the PDF is obtained empirically.

4.1. Qualitative Evaluation of PDF Predictions

In the simplest case, we only have one model with one set of model parameters to cover all predictions. In this case, the evaluation of the full probability distribution is straight-forward: We would fill a histogram of all observed values, such as sales records, and overlay this with the single model, such as a negative binomial with predicted parameters, that is used for all observations. Then, we compare the model curve directly with the observations, using statistical tests such as the Kolmogorov-Smirnov test.

In practical applications however, we have a large number of effective prediction models, since although we always use the same model parameterization, such as the negative binomial distribution, its parameters have to be determined at the required level of granularity. For example, for daily orders, we need to predict the parameters of the negative binomial distribution for each location, sales day, and product. Unlike the simple case discussed above, where we had many observations to compare the prediction model to, we now have just a single observation per prediction, meaning that we cannot use statistical tests directly.

4.1.1. Histogram of CDF Observations

For a first qualitative assessment, we make use of the probability integral transform, see e.g. [33, 34], which states that a random variable distributed according to the CDF of another random variable is uniformly distributed. So, in case of a correctly calibrated PDF prediction, the distribution of the corresponding actually observed CDF values of the individual PDF predictions is expected to be uniform, *regardless* of the shape of the predicted distribution, and any deviation can be interpreted as a hint that the predicted PDF is not fully correct [35].

The CDF of a PDF $f(x)$ is defined as:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(x')dx' \tag{8}$$

Here, $F_X(x)$ is the CDF with $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_x(x) = 1$. The cumulative distribution describes the probability that a the variable has a value smaller than x and intuitively represents the area under $f(x')$ up to a point x.

If the CDF is continuous and strictly increasing, then the inverse of the CDF, $F^{-1}(y)$, exists and is a unique real-valued number x for each $y \in [0, 1]$, so that we can write $F(x) = y$. The latter defines the inverse quantile function, because we can define the quantile τ of the probability distribution $f(x)$ as:

$$Q_\tau = F^{-1}(\tau) \tag{9}$$

Using the example of the normal distribution with $\mathcal{N}(0, 1)$ as shown in Fig. 1, we can identify the median ($\tau = 0.5$) by first looking at the CDF in the lower part of the figure, look at $y = 0.5$ on the y-axis and then identify the point on the x axis for both the PDF $f(x)$ and the CDF $F(x)$ that correspond to the quantile τ. In the case of the normal distribution, this is of course the central value at zero.

We can then interpret the CDF as a new variable $s = F(t)$, meaning that F becomes a transformation that maps t to s, i.e. $F : t \to s$. Accordingly, $\lim_{t\to-\infty} s(t) = 0$ and $\lim_{t\to\infty} s(t) = 1$ and s can be intuitively interpreted as the fraction of the distribution of t with values smaller than t from the definition of the CDF. This implies that the probability distribution of s, $g(s)$, is constant in the interval $s \in [0, 1]$ in which s is defined, and s can be interpreted as the cumulative distribution of its own probability distribution:

$$s = G(s) = \int_{-\infty}^{s} g(s')ds' \tag{10}$$

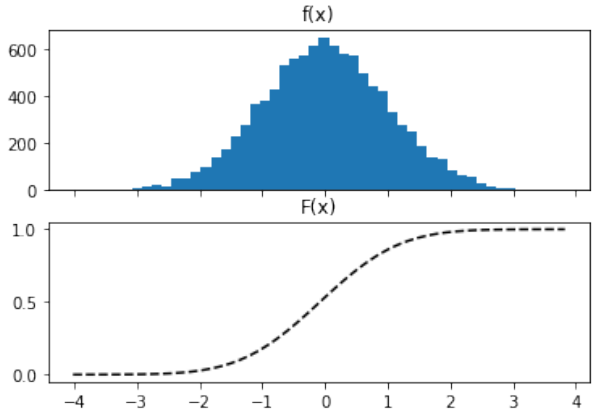

Figure 1: Probability distribution function and cumulative distribution function of a normal distribution.

In case of discrete probability functions, such as the negative binomial function, the same argument still holds, but the definition of the inverse quantile function is replaced by the generalized inverse: $F^{-1}(y) = \inf\{x : F(x) > y\}$ for $y \in [0, 1]$, see e.g. [34, p. 54]. In order to obtain a uniform distribution for discrete PDFs that is comparable to the case of continuous distributions, the histogram holding the inverse quantiles or values of the CDF is filled using random numbers according to the intervals of the CDF. For example, if the sales of zero items accounts for 75% of the observed sales distribution for this item, the value of the CDF function that is used to fill the histogram is randomly chosen in the interval $[0, 0.75]$. Proceeding similarly for all other observed values, the resulting histogram of CDF values should again be uniform as in the case of a continuous PDF.

A histogram of the observed quantiles (or CDF values) for each individual PDF prediction is therefore expected to be uniformly distributed in $[0, 1]$, if the predicted probability distribution $f(x)$ is correctly calibrated. This is illustrated in Fig. 2, which shows the distribution of observed quantiles for five different cases. If both the choice of the model and the model parameters are estimated correctly, we would expect the uniform distribution. If the mean or the variance are not estimated correctly, the resulting distribution will show a distinct deviation from this uniform behavior.

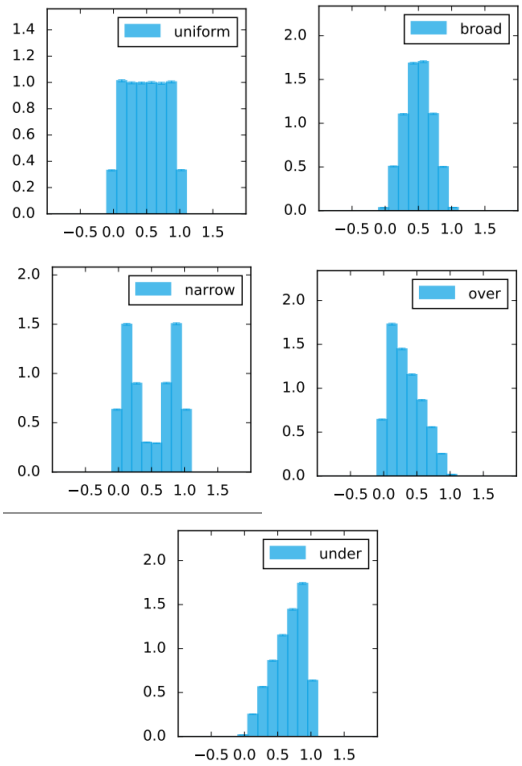

Figure 2: Quantile distribution for different cases of estimating the full probability distribution compared to the expected distribution: correct prediction ("uniform"), variance overestimated ("broad"), variance underestimated ("narrow"), mean overestimated ("over"), mean underestimated ("under")

4.1.2. Quantile Profile Plot

We now refine the method from sec. 4.1.1 by comparing (in the sense of higher or lower) the quantiles of the observed events, i.e. the sales records, with specified quantile values. In order to do so, we start again from the predicted values for the mean and variance from which a negative binomial distribution is constructed for each prediction, for example for the predicted demand for a single product sold on a single day in a given sales location. Each of these predicted negative binomial PDFs is then transformed to its CDF. Note that for simplicity, we always refer to the negative binomial model in this description, however, the general approach is valid for any PDF.

Then we compare the actual observed sales value (corrected for censored data if necessary) to different quantiles of the corresponding predicted distribution for each data record and average over a larger data sample. For example, if we wanted to check that the median of the distribution, corresponding to the 0.5 quantile, is predicted correctly by the machine learning model, we would compare the quantile value 0.5 to the ratio of quantile values of observed sales records being lower/higher than 0.5. In other words, in case of the median, 50% of the ex post observed target values should be observed below the median of the corresponding individual predicted PDF and 50% above.

In order to judge whether the overall shape of the predicted distributions is predicted correctly, we repeat this procedure for a range of quantiles, for example $q = 0.1, 0.2, \ldots, 0.9$. However, we are free to choose which quantiles to look at, and in specific situations it might be advisable to look at the tails of the distribution in more detail, to make sure that even relatively rare events are estimated correctly by the machine learning algorithm, and add more quantiles for comparison in the region between, say $q = 0.95$ and $q = 0.99$. In the following, we call this method *quantile profile plot*.

Fig. 3 illustrates five different collections of quantile profile plots (each collection comparing to 7 specified quantile values), for separate sets of exemplary PDF estimations and observed data combinations. The dashed horizontal lines indicates the fraction we expect, i.e. the specified quantile value, if the predictions are correct. For example, for the median, the line at 0.5 indicates that 50 percent of all PDF prediction and observed data combinations in a given data set should fall above the line, and 50 percent should fall below the line. The observation of the number of samples, indicated with shaded circles, that do in fact fall above and below a particular line, then allows the evaluation of the accuracy of PDF estimations. In case that the PDFs are not estimated correctly, the fractions will deviate from their expected values and the corresponding profile plot allows to judge whether for example the tails of the predicted distribution describing rare events are particularly problematic.

In the same way as the method of filling all observed CDF values of the individual PDF predictions in a histogram and compare to a uniform distribution (described in sec. 4.1.1), quantile profile plots do not work on individual PDF predictions but require a statistical population. However, calculating the fractions of the quantile profile plots globally, i.e. over all samples in the data set, might not reveal certain deficiencies for a subset, e.g. a specific store or product group in the example of retail demand forecasts. Therefore, we combine the approach described above with the method of profile plots, where the quantity on the x-axis of the quantile profile plot can be any variable of the data set at hand. Profile plots are akin to scatter plots and described in more detail in appendix A.

In summary, this approach has two major benefits compared to the method discussed in sec. 4.1.1: First, by explicitly visualizing several different quantiles, the quantile profile plot reveals which part of the predicted PDF, such as the tails of the probability distribution, are particularly problematic. Second, by showing the dependency

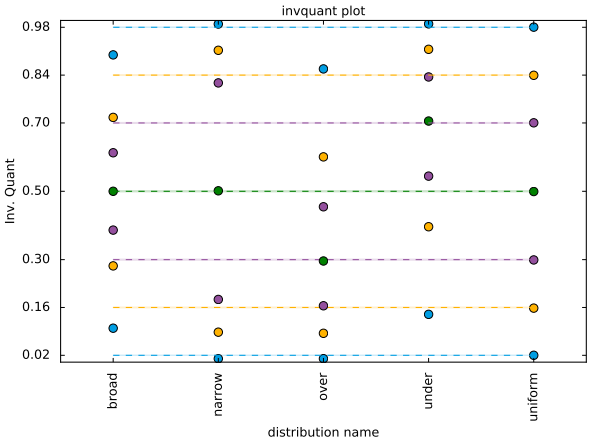

Figure 3: Profile plots of quantiles comparing all predicted PDFs to the corresponding observed events. In the leftmost two columns of shaded circles illustrate the behavior if the estimate of the variance is biased, the center and center-right columns of shaded circles illustrate cases for which the mean estimation is biased and the far right column shows the expected behaviour if all predictions are correct.

from the (arbitrary) variable on the x-axis, the quantile profile plot reveals deviations of the predicted PDF from the actuals for different parts, e.g. specific categories, of that variable. Two examples for this can be found in figures 7 and 8 in the next section.

4.2. Quantitative Evaluation of PDF Predictions

The methods discussed so far allow a detailed qualitative evaluation of PDF predictions. However, in order to also quantify the quality of the PDF predictions, a metric is required that assesses the difference between the distributions. Therefore, we want to compare the CDF histogram of the predicted PDFs with the expected uniform distribution and define a metric in the range between 0 and 1, such that the metric takes the value of 1 when both distributions agree perfectly. Several approaches that measure the difference between two probability distributions are suggested in the literature such as the first Wasserstein distance [36], the Kullback-Leibler divergence [37], and the Jensen-Shannon divergence [38].

The first Wasserstein distance, also known as earth mover distance (EMD), represents a distance defined between two probability distributions on a given metric space, and for our purposes here can defined by:

$$\mathrm{EMD}(P, Q) = 2 \cdot \frac{\sum_{k=1}^{N} |F_P(x_k) - F_Q(x_k)|}{N}, \qquad (11)$$

where $F_P(X)$ and $F_Q(X)$ are the CDFs of the two PDFs $P(X)$ and $Q(X)$, respectively, and x_k denotes the average value of X in bin k, with X being divided in N bins. Compared to the other metrics given below, the first Wasserstein or earth mover distance has the additional benefit,

that it takes a specific metric space into account, meaning that it depends on the distance of potential deviations.

The Kullback-Leibler divergence is a measure of how one probability distribution diverges from a second expected probability distribution. For PDFs $P(X)$ and $Q(X)$, again with X being divided in N bins k, the Kullback-Leibler divergence from Q to P is defined as:

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{k=1}^{N} P(x_k) \log \left(\frac{P(x_k)}{Q(x_k)} \right), \qquad (12)$$

where the logarithm can be either base-2 or the natural logarithm.

The Jensen-Shannon divergence (JSD) can be seen as a symmetrized and smoothed version of the Kullback-Leibler divergence:

$$D_{\mathrm{JSD}}(P \parallel Q) = 0.5 \cdot (D_{\mathrm{KL}}(P \parallel M) + D_{\mathrm{KL}}(P \parallel M)), \quad (13)$$

where $M = 0.5 \cdot (P + Q)$.

5. Example: Demand Forecasting

We use data from a Kaggle online competition about estimating unit sales of Walmart retail goods [39] to demonstrate forecasting of demand for different product-location-date combinations in form of full individual PDFs by means of Cyclic Boosting and subsequent qualitative and quantitative evaluation of the PDF predictions. The fields identifying a unique sample of the data set are store (`store_id`) and item identifiers (`item_id`) as well as date. The target y which we need to predict is the number of sales of a given product in a given store on a specific day, denoted by `sales`. In the following, we use data from 2013-01-01 to 2016-05-22, that describe the sales of 100 different products (`FOODS_3_500`, ..., `FOODS_3_599`) of the department `FOODS_3` in 10 stores. All data before 2016 are used as the training data and the data from 2016 are used as an independent test or validation set. Besides the fields used to identify an individual sales record and the corresponding observed sales value, namely `item_id`, `store_id`, `date`, `sales`, we also use the fields `event_name_1`, `event_type_1`, `snap_CA`, `snap_TX`, `snap_WI`, `sell_price`, and `list_price` to build features for our machine learning models.

5.1. Mean Estimation

The first model, based on the Cyclic Boosting approach described above, is used to predict the mean of an assumed negative binomial distribution, and we use the following variables as features: categorical variables for `store_id` and `item_id`, several derived variables that are constructed from the time-series of the sales records describing trend and seasonality (days since beginning of 2013 as linear trend as well as day of week, day of year, month, and week of month), time windows around the events given in the data set (7 days before until 3 days after for Christmas and Easter, and 3 days before until 1 day after for all other events like New Year or Thanksgiving), a flag denoting a promotion, and the ratio of reduced (`sell_price`) and normal price (`list_price`). We also include various

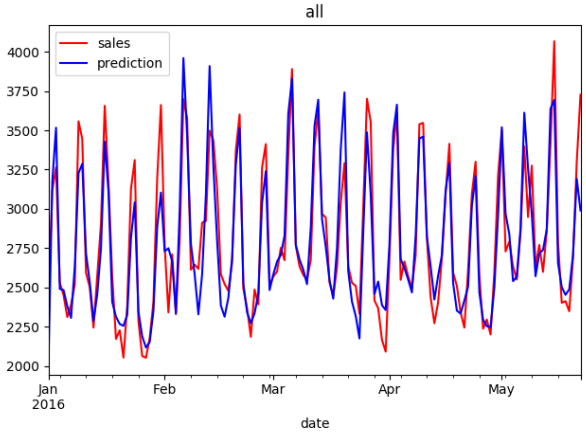

Figure 4: Time series of mean prediction and sales in test period summed over all products and stores.

two-dimensional combinations of these features. In these cases, one of the two dimensions is either `store_id` or `item_id`, allowing the machine learning model to learn characteristics of individual locations and products.

Unlike most state-of-the-art time series forecasting methods, we do not include lagged target information, for example via stacking of exponentially weighted moving average features, in our model. The reason for this is that including such information makes the learning of exogenous effects, e.g. promotions or events, much harder, as the model tends to rely mainly on temporal confounding. Omitting these kind of variables improves the capability of the machine learning model to learn causal dependencies, which in turn improves the explainability of the model as well as the quality of the forecasts for mid- to long-term predictions and rare events. In order to capture recent trends that are not reflected in the exogenous features of the model, we apply an individual residual correction on each of the predictions of the machine learning model, which accounts for deviations between the exponentially weighted moving average (with a recursive smoothing factor of 0.15) of the predictions and targets of each product-location combination over the corresponding past. To reflect a realistic replenishment scenario, we use a prediction horizon of two days for our example here, i.e. a target lag of two days for the training.

Just to give a rough verification for the accuracy of the mean predictions, we report the MAD and MSE averaged over the full used data set in the test period to be 1.65 and 10.15, respectively, while the average of the target, i.e. the sales, is 3.28. Fig. 4 shows the time series of both predictions and sales summed over all 100 products and 10 stores during the test period.

5.2. Variance Estimation

The second model, also based on Cyclic Boosting, is used to estimate the variance of the negative binomial distribution. The data are split into training and test set as above and in addition the mean predictions for each individual product-location-date combination are fixed in the variance model

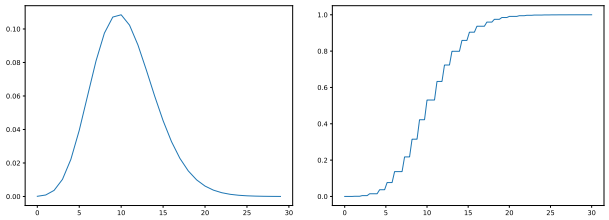

Figure 5: Predicted PDF (left) and CDF (right) for a specific product-location-date combination.

as stated by Eqn. (6). This effectively means that the mean predictions are created in-sample for the training period using the fully trained and validated model for the mean discussed above.

In this model focusing on the variance, we use the same set of features as for the mean model described above in sec. 5.1, except for dropping the two-dimensional combinations including individual event features. An example for the resulting PDF and CDF predictions for item `FOODS_3_586` in store `WI_1` on 2016-05-22 is shown in Fig. 5.

5.3. Evaluation of PDF Predictions

Fig. 6 shows the histogram of CDF observations according to the method described in sec. 4.1.1 for all product-location-day combinations in the test period. As benchmark, we compare the outcome of our negative binomial model to a simpler Poisson assumption, which has only a single model parameter, the mean. Using the same mean predictions for both negative binomial and Poisson model, the negative binomial PDF predictions are much closer to the uniform distribution, which we expect for optimal PDF predictions, than the Poisson PDF predictions, showing the effectiveness of our variance estimation. The only significant deviation of the negative binomial histogram from the uniform distribution can be found in the last bins of CDF values close to 1. As can be seen when excluding samples with mean predictions lower than 1.0, this deviation stems from slow-sellers, which are prone to overdispersion.

Fig. 7 and Fig. 8 show quantile profile plots for different variables on the x-axis, namely mean predictions and day of week, respectively. All product-location-day combinations in the test period are used to generate the statistics in the plots. It can be seen that the different weekdays (from Monday as 0 to Sunday as 6) show slightly different patterns and that there are larger deviations for higher mean predictions (with fewer samples though).

The quantitative results for the CDF accuracies of our PDF predictions for the different metrics described in sec. 4.2, calculated over all product-location-day combinations in the test period, can be found in Tab. 1. As benchmark, we compare these to the results of a Poisson model assumption needing only the mean as single parameter. Our negative binomial PDF predictions show a significant improvement over the simpler Poisson model, which uses the same mean predictions, for each of the metrics.

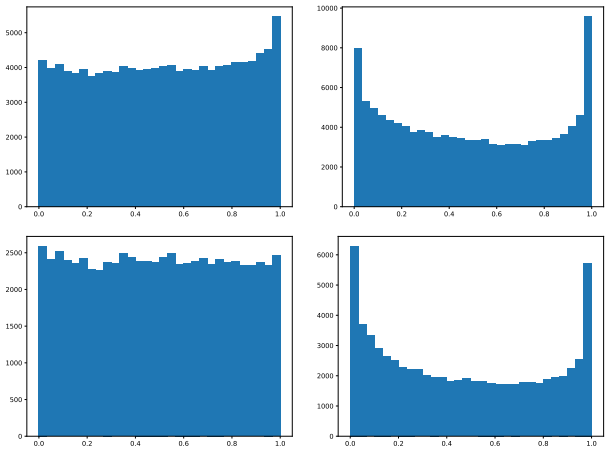

Figure 6: Histograms of ex post observed target CDF values of the corresponding individual PDF predictions (to be compared to a uniform distribution) for our negative binomial model (left) and a simpler Poisson model for comparison (right), using all product-location-day combinations in the test period (upper two plots). In order to show the effect of slow-sellers, the lower two plots exclude all samples with mean predictions lower than 1.0.

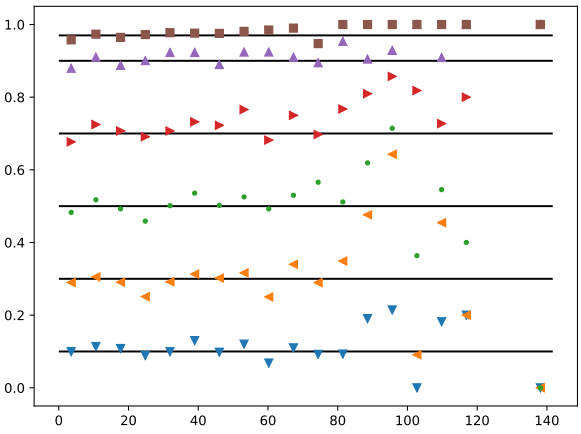

Figure 7: Quantile profile plot for mean predictions on the x-axis aggregated over all product-location-day combinations in the test period.

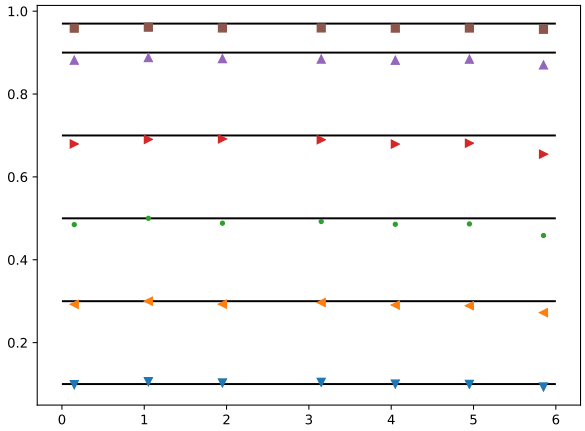

Figure 8: Quantile profile plot for the different days of week on the x-axis (from Monday to Sunday) aggregated over all product-location-day combinations in the test period.

Table 1: CDF accuracies for negative binomial and Poisson PDF predictions for different metrics calculated over all product-location-day combinations in the test period.

	NBD	**Poisson**
EMD	0.965	0.852
KL_2	0.995	0.933
KL_e	0.997	0.954
JSD_2	0.999	0.982
JSD_e	0.999	0.987

6. Conclusion

We have shown how to use two subsequent machine learning models for mean and variance estimation together with a negative binomial model assumption to come up with individual PDF predictions. This setup is especially useful for retail demand forecasting, where the true demand follows a negative binomial distribution quite closely. Compared to a model-free approach like quantile regression, the distributional assumption therefore drastically reduces the uncertainty of the resulting predictions, which are, by using Cyclic Boosting as underlying machine learning algorithm, fully explainable on the individual level.

Furthermore, we have presented new qualitative and quantitative methods for evaluating predictions in form of full PDFs. For qualitative evaluation, it is important to check the full PDF, especially its tails that are often crucial for subsequent decision making, as well as investigate aggregations of individual PDFs (to reduce the uncertainty of the evaluation) dependent on specific variables, and we have proposed a novel profile approach called quantile profile plot for this. For quantitative evaluation, it is always desirable to have a single number to compare different models in terms of accuracy, and we have suggested to use the deviance between the CDF histogram of the predicted PDFs and the uniform distribution to compare PDF accuracies.

References

[1] A.-L. Beutel and S. Minner, "Safety stock planning under causal demand forecasting," *International Journal of Production Economics*, vol. 140, no. 2, pp. 637–645, 2012.

[2] G.-Y. Ban and C. Rudin, "The big data newsvendor: Practical insights from machine learning," *Operations Research*, vol. 67, no. 1, pp. 90–108, 2019.

[3] D. Bertsimas and N. Kallus, "From predictive to prescriptive analytics," *Management Science*, vol. 66, no. 3, pp. 1025–1044, 2020.

[4] A. Oroojlooyjadid, L. V. Snyder, and M. Takáč, "Applying deep learning to the newsvendor problem," *IISE Transactions*, vol. 52, no. 4, pp. 444–463, 2020.

[5] J. Huber, S. Müller, M. Fleischmann, and H. Stuckenschmidt, "A data-driven newsvendor problem: From data to decision," *European Journal of Operational Research*, vol. 278, no. 3, pp. 904–915, 2019.

[6] J. Pearl, *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2 ed., 2009.

[7] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies.," *Journal of educational Psychology*, vol. 66, no. 5, p. 688, 1974.

[8] E. A. Silver, D. F. Pyke, R. Peterson, *et al.*, *Inventory management and production planning and scheduling*, vol. 3. Wiley New York, 1998.

[9] H. Scarf, *A min-max solution of an inventory problem*. Studies in The Mathematical Theory of Inventory and Production, Stanford University Press, 1958.

[10] H. Scarf, "The optimality of (s,s) policies in the dynamic inventory problem," *Mathematical Methods in the Social Sciences*, 1959.

[11] S. Nahmias and W. P. Pierskalla, "Optimal ordering policies for a product that perishes in two periods subject to stochastic demand," *Naval Research Logistics Quarterly*, vol. 20, no. 2, pp. 207–229, 1973.

[12] S. Nahmias, "Optimal ordering policies for perishable inventoryii," *Operations Research*, vol. 23, no. 4, pp. 735–749, 1975.

[13] S. Nahmias, "The fixed-charge perishable inventory problem," *Operations Research*, vol. 26, no. 3, pp. 464–481, 1978.

[14] S. Minner and S. Transchel, "Periodic review inventory-control for perishable products under service-level constraints," *OR spectrum*, vol. 32, no. 4, pp. 979–996, 2010.

[15] F. Edgeworth, "The mathematical theory of banking," *Journal of the Royal Statistical Society*, 1888.

[16] M. Khouja, "The single-period (news-vendor) problem: literature review and suggestions for future research," *Omega*, vol. 27, no. 5, pp. 537 – 553, 1999.

[17] L. C. Alwan, M. Xu, D.-Q. Yao, and X. Yue, "The dynamic newsvendor model with correlated demand," *Decision Sciences*, vol. 47, no. 1, pp. 11–30, 2016.

[18] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on stochastic programming: modeling and theory*. SIAM, 2014.

[19] S. Saghafian and B. Tomlin, "The newsvendor under demand ambiguity: Combining data with moment and tail information," *Operations Research*, vol. 64, no. 1, pp. 167–185, 2016.

[20] R. Koenker and K. F. Hallock, "Quantile regression," *Journal of economic perspectives*, vol. 15, no. 4, pp. 143–156, 2001.

[21] R. Wen, K. Torkkola, and B. Narayanaswamy, "A multi-horizon quantile recurrent forecaster," 11 2017.

[22] M. Feindt and U. Kerzel, "The neurobayes neural network package," *NIM A*, vol. 559, no. 1, pp. 190 – 194, 2006.

[23] C. M. Bishop, "Mixture density networks." 1994.

[24] D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski, "Deepar: Probabilistic forecasting with autoregressive recurrent networks," *International Journal of Forecasting*, vol. 36, no. 3, pp. 1181–1191, 2020.

[25] I. Adan, M. van Eenige, and J. Resing, "Fitting discrete distributions on the first two moments," *Probability in the engineering and informational sciences*, vol. 9, no. 4, pp. 623–632, 1995.

[26] A. S. C. Ehrenberg, "The pattern of consumer purchases.," *Journal of the Royal Statistical Society Series C*, no. 1, p. 26, 1959.

[27] G. J. Goodhardt and A. Ehrenberg, "Conditional trend analysis: A breakdown by initial purchasing level," *Journal of Marketing Research*, vol. IV, pp. 155–161, May 1967.

[28] A. Ehrenberg, *Repeat-buying; theory and applications*. North-Holland Pub. Co., 1972.

[29] C. Chatfield and G. J. Goodhardt, "A consumer purchasing model with erlang inter-purchase time," *Journal of the American Statistical Association*, vol. 68, no. 344, pp. 828–835, 1973.

[30] D. C. Schmittlein, A. C. Bemmaor, and D. G. Morrison, "Technical notewhy does the nbd model work? robustness in representing product purchases, brand purchases and imperfectly recorded purchases," *Marketing Science*, vol. 4, no. 3, pp. 255–266, 1985.

[31] F. Wick, U. Kerzel, and M. Feindt, "Cyclic boosting - an explainable supervised machine learning algorithm," in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pp. 358–363, IEEE, Dec. 2019.

[32] J. Hilbe, *Negative binomial regression*. Cambridge, UK New York: Cambridge University Press, 2011.

[33] J. E. Angus, "The probability integral transform and related results," *SIAM Review*, vol. 36, no. 4, pp. 652–654, 1994.

[34] G. Casella, *Statistical inference*. Pacific Grove, Calif: Duxbury/Thomson Learning, 2002.

[35] F. X. Diebold, T. A. Gunther, and A. S. Tay, "Vevaluating density forecasts with applications to financial risk management. v inter# national economic review, vol. 39, no. 4," in *Symposium on Forecasting and Empirical Methods in Macroeconomics and Finance*, p. 863, 1998.

[36] I. Olkin and F. Pukelsheim, "The distance between two random vectors with given dispersion matrices," *Linear Algebra Appl.*, vol. 48, pp. 257–263, 1982.

[37] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, pp. 79–86, 03 1951.

[38] I. Dagan, L. Lee, and F. Pereira, "Similarity-based methods for word sense disambiguation," in *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, ACL '98/EACL '98, (USA), pp. 56–63, Association for Computational Linguistics, 1997.

[39] `https://www.kaggle.com/c/m5-forecasting-accuracy/data`.

A. Profile Histograms

In many cases, scatter plots are used to study the behaviour of two distributions or sets of data points visually. However, even for moderate amount of data, this approach becomes quickly difficult. To illustrate this, a sample of (x, y) data points was obtained in the following way: The distribution of x values was obtained by generating 5,000 samples of Gaussian distributed random numbers $X \sim \mathcal{N}(0.0, 2.0)$ and the y values are obtained via $Y \sim X + \mathcal{N}(2.0, 1.5)$. Fig. 9 shows the marginal distributions for x and y as well as a scatter plot of x vs. y.

Although the simple linear correlation between X and Y is apparent in the scatter plot, finer details are not visible and it is easy to imagine that a more complex relationship is difficult to discern. Profile histograms are specifically designed to address this shortcoming. Intuitively, profile histograms are a one-dimensional representation of the two-dimensional scatter plot and are obtained in the following way: The variable on the x axis is discretized into a suitable

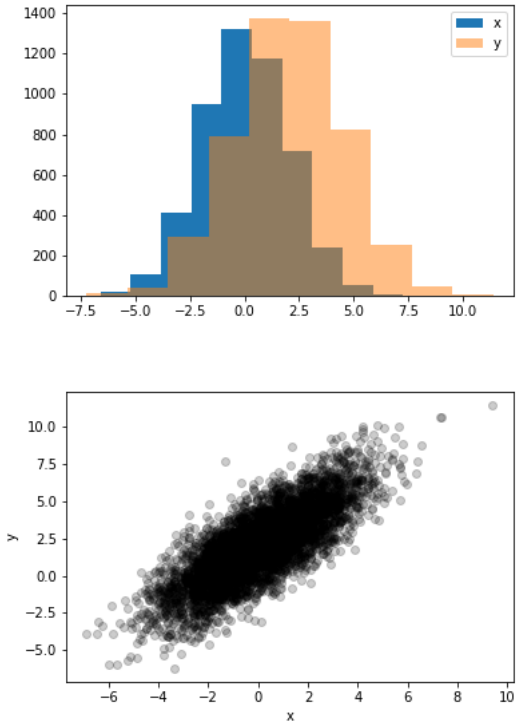

Figure 9: Marginal distribution and scatter plot of variables X and Y.

range of bins. The exact choice of binning depends on the problem at hand. One can for example choose equidistant bins in the range of the x axis or non-equidistant bins such that each bin contains the same number of observations. Then within each bin of the variable X, the a location and dispersion metric is calculated for the variable Y. This means that the bin-borders on the X axis are used as constraints on the variable Y and with these conditions applied, for example the sample mean of the selected y values as well as the standard deviation are calculated. These location and dispersion metrics in each bin of X are used to illustrate the behaviour of the variable Y as the values of the variable X change from bin to bin. The resulting profile histogram is shown in Fig. 10. This one-dimensional representation allows to understand even a complex relationship between two variables visually. Note that due to few data points at the edges of the distributions the profile histogram is expected to show visual artifacts in the corresponding regions.

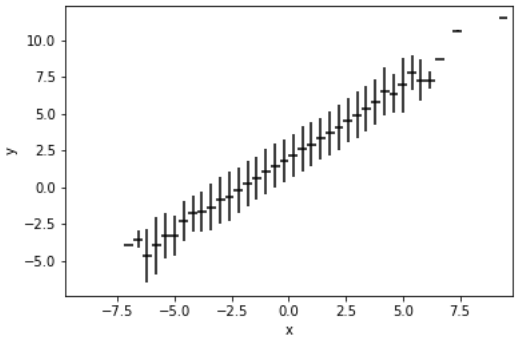

Figure 10: Profile histogram of variables X and Y.