Research Collection School Of Information Systems

School of Information Systems

4-2020

DFSeer: A visual analytics approach to facilitate model selection for demand forecasting

Dong SUN

Zezheng FENG

Yuanzhe CHEN

Yong WANG
Singapore Management University, yongwang@smu.edu.sg

Jia ZENG

See next page for additional authors

Citation

1

Author

Dong SUN, Zezheng FENG, Yuanzhe CHEN, Yong WANG, Jia ZENG, Mingxuan YUAN, Ting-Chuen PONG, and Huamin QU

DFSeer: A Visual Analytics Approach to Facilitate Model Selection for Demand Forecasting

Dong Sun[1], **Zezheng Feng**[1], **Yuanzhe Chen**[2], **Yong Wang**[1], **Jia Zeng**[2], **Mingxuan Yuan**[2],
Ting-Chuen Pong[1], **Huamin Qu**[1]

[1]The Hong Kong University of Science and Technology
[2]Noah's Ark Lab, Huawei Technologies Co. Ltd.
[1]{dsunae, zfengak, ywangct, tcpong, huamin}@cse.ust.hk; [2]{chenyuanzhe, zeng.jia,
yuan.mingxuan}@huawei.com

ABSTRACT

Selecting an appropriate model to forecast product demand is critical to the manufacturing industry. However, due to the data complexity, market uncertainty and users' demanding requirements for the model, it is challenging for demand analysts to select a proper model. Although existing model selection methods can reduce the manual burden to some extent, they often fail to present model performance details on individual products and reveal the potential risk of the selected model. This paper presents DFSeer, an interactive visualization system to conduct reliable model selection for demand forecasting based on the products with similar historical demand. It supports model comparison and selection with different levels of details. Besides, it shows the difference in model performance on similar products to reveal the risk of model selection and increase users' confidence in choosing a forecasting model. Two case studies and interviews with domain experts demonstrate the effectiveness and usability of DFSeer.

Author Keywords

Interactive visualization; model selection; product demand forecasting; time series.

CCS Concepts

•**Human-centered computing** → **Visualization; User interface design; Human computer interaction (HCI);**

INTRODUCTION

In a manufacturing company, the forecast of product demand based on the historical demand data is critical to production planning and supply strategy developing [29]. Since there is often no model that is suitable for forecasting all the product demand [1, 22], it has become increasingly important to select appropriate models for demand forecasting. An appropriate model for demand forecasting should be accurate, stable, and

suitable for multiple products, which can optimize the allocation of production resources and reduce the risk of both order delay and the overstock of products [29].

However, it is challenging to select a proper model for demand forecasting. First, it is not easy to evaluate and compare different demand forecasting models due to the data complexity. The demand forecasting in a manufacturing company usually involves thousands of products and dozens of models. The domain experts often find it difficult to quickly evaluate and compare different demand forecasting models when the model performance is inconsistent on different products or in different time periods for the same product. Second, the market uncertainty, which is common in the manufacturing industry, makes it even more difficult to find a reliable demand forecasting model [39]. For example, a sudden decline in the price of substitute products may lead to a reduction in the demand, which can result in a large difference between the forecasted and exact demand. Third, it is challenging to find a model that fulfills the high requirements of product demand forecasting. The demand analyst is seeking a model that is accurate for most of the products, stable in the accuracy when forecasting the demand in different months, and applicable to as many products as possible.

Some prior research has been conducted on the selection of forecasting models, which mainly focuses on parameter tuning [6], feature selection for modeling [36, 42], and the comparison of different types of models [23, 30, 45, 52, 56, 40]. These studies provide helpful guidance for model selection and reduce lots of manual efforts in model selection. However, they often only provide the overall performance of models on the dataset and fail to reveal the performance details on individual products, which is important for decision making in demand forecasting model selection. For example, *how accurate is the product demand forecast by different models? How different is the demand forecasting performance of a certain model on different types of products (or different products of the same type)? Are there some models that are suitable for the demand forecasting of most products?* In addition, these studies do not reveal the risk of using a specific demand forecasting model. Instead of only getting a ranked list of models, users are also curious about the reliability of the models in forecasting the demand of the products with similar historical demand data,

which is crucial to their choice of an appropriate demand forecasting model [10, 4]. Employing a high-risk model can lead to a reduction in revenue, loss of customers, and users' distrust of the model.

In this paper, we propose DFSeer, an interactive visual analytics system to support the comparison and selection of demand forecasting models in the manufacturing industry. Considering that a model with a good average performance on all the products may produce a poor forecast for some products, DFSeer supports model selection according to the forecast of the products with similar historical demand data [10, 4, 21]. Our approach allows users to compare and select demand forecasting models through three levels of details, i.e., an overview, a product view, and a detail view. It can help users easily identify the models whose forecasting results are inaccurate or unstable for some products, as well as the models that are only suitable for a small number of products. Besides, the system combines the model performance with the historical demand data to reveal the risk of selecting a specific demand forecasting model. It allows users to remove dissimilar products identified by integrating their domain knowledge with the properties of the time series extracted by automated algorithms. By showing the difference in model performance on the products with similar historical demand, DFSeer can help users understand the possibility of an accurate forecast by the selected model, which can increase their confidence in their choice of the model. To the best of our knowledge, DFSeer is the first visual analytics system to support the interactive selection of different types of time series forecasting models.

We summarize the major contributions of the work as follows:

- A new workflow that supports demand forecasting model selection based on the products with similar historical demand and reveals the risk of selecting a model to increase users' confidence in adopting the model.
- The implementation of a visual analytics system that enables users to compare and select demand forecasting models with different levels of details.
- Detailed evaluations of the visualization system, including two case studies with real-world product demand data and interviews with four domain experts from a manufacturing company.

RELATED WORK
Previous studies related to our work can be divided into three parts: techniques for time series forecasting, model selection for forecasting analysis, and visualization of discrete time series.

Techniques for Time Series Forecasting
Recently a number of techniques have been proposed and applied to the forecast of time series, which can be categorized into three groups according to previous studies [20, 49]: classical techniques for time series modeling, machine learning techniques, and deep learning techniques.

Classical techniques for time series modeling mainly includes the autoregressive integrated moving average (ARIMA) model [8] and the exponential smoothing (ES) model [33]. The

ARIMA model [8] computes the difference between two adjacent values in a time series to convert non-stationary data to stationary ones and utilizes the accumulated error terms to reduce the random fluctuation in time series forecasting. The ES model [33] emphasizes the effect of the recent time series on the forecast, without ignoring the effect of the time series from a long time ago. Machine learning techniques have also been applied to time series forecasting. For example, the k-nearest neighbor regression (KNN) model [3] forecasts the future value of a time series based on its neighboring time series. Other machine learning techniques include the support vector regression (SVR) model [28], gradient boosting [25], Lasso [53], ridge regression [31] and so on. In addition, inspired by the excellent performance of deep learning techniques [38], researchers have also introduced deep learning techniques to time series forecasting. For instance, Model-Agnostic Meta-Learning (MAML) [24] can be trained quickly even though there are only a small number of data points. When forecasting the future values of a time series, the recurrent neural network (RNN) [43] takes both the current value and past ones into consideration.

Each of these techniques performs well on certain forecasting tasks, but their performance often declines on other forecasting tasks [1]. In this paper, we build an interactive visual analytics system to facilitate the fast selection of demand forecasting models and help users easily identify the risk of a model.

Model Selection for Forecasting Analysis
Various studies have been conducted to facilitate the selection of forecasting analysis models. These studies mainly consist of two types: visualization of forecasting models and automated approaches for model selection.

Visual analytics techniques play an important role in helping users understand forecasting models. For example, Krause et al. [37] and Hohman et al. [32] studied the influence of a feature on the forecast by varying the value of the feature. Visual analytics approaches have also been employed to explain the mechanism of deep learning models [41, 44, 55]. Some other studies focus on using visualization techniques to facilitate the comparison and selection of forecasting models. For instance, TiMoVA [6] provides a detailed comparison of ARIMA models with different parameter settings to seek the best model. INFUSE [36] visualizes the significance of various features in forecasting analysis with a glyph design, where users can select different features to train a model and compare the generated models. ATMSeer [56] exploits visualization techniques to support the interactive exploration of a model selection process. Techniques in the statistics literature, such as the residual plot [16] and the quantile-quantile (Q-Q) plot [59], have also been employed to evaluate and select models. The worm plot [11] supports parameter tuning and the selection of LMS models [15]. Gabry et al. [26] introduced visualization techniques to the building of Bayesian models to help model developers evaluate and compare models. Apart from these visualization methods, some automated approaches for model selection have also been proposed. For example, Auto-WEKA 2.0 [35] and H2O AutoML [30] are developed to help data scientists tune the parameters of different algorithms, compare

the explored models, and find the model that performs the best on the dataset specified by the user.

Although these techniques provide useful guidance for model selection and reduce the workload to some extent, they are not applicable to model selection for demand forecasting for two reasons. First, these approaches are aimed at the optimized performance of the whole dataset, which may result in an inaccurate forecast of the demand for some products. Second, they lack the performance details of models on individual products, which is important to decision making in demand forecasting model selection. Our work is proposed to solve the aforementioned problems, which supports demand forecasting model selection with different levels of details based on the forecast of the products with similar historical demand. Besides, it reveals the risk of the selected model by showing the performance difference on similar products.

Visualization of Discrete Time Series
Visual analytics approaches are widely used to explore time series [46, 5]. According to the survey by Aigner et al. [2], the studies on time series data visualization mainly includes three parts: ordinal, discrete, and continuous time series visualization. We will only discuss the prior work on visualizing discrete time series, which is more related to our approach.

One of the most popular visualization techniques for discrete time series is the line chart [47], which describes how data change over time by connected line segments. An alternative of the line chart is the horizon graph [48], a space-saving technique that encodes the data by both the color and the height. Besides, the spiral [58] and calendar-based visualization [54] can reveal the periodic features of time series. To support the exploration of large scale time series, plenty of visual analytics tools have been developed. For example, PlanningVis [51] is a multi-level visualization system that juxtaposes heatmaps, line charts and bar charts to assist factory planners in optimizing a production plan.

Inspired by the prior studies, we propose novel visualization techniques to show the discrete time series data and support the selection of demand forecasting models.

SYSTEM REQUIREMENTS AND DESIGN
In this section, we interviewed four domain experts to understand the problem of model selection for product demand forecasting. Based on the comments and suggestions collected from the interviews, we summarized five requirements to guide the design of the visualization system.

Interview Study
In order to understand the workflow of model selection for product demand forecasting, identify existing challenges in the approach currently used, and seek opportunities to improve the process of model selection, we conducted interviews with four domain experts from a manufacturing company. Two of the domain experts are algorithm developers who apply and train various models to forecast the demand of thousands of products, while the other two are demand analysts who are required to compare the performance of the models provided by algorithm developers and select the most suitable model.

The interview for each domain expert lasted for one hour, which can be divided into three parts: the introduction to demand forecasting, the description of the workflow of model selection, and the discussion of challenges and user requirements. In the first part, the domain expert introduced the demand forecasting problem, the demand data used, and the algorithms to forecast product demand. In the second part, we asked the domain expert questions about the workflow of model selection for demand forecasting, including *what features will be used to evaluate a model, what weights will be assigned to these features to rank the models, what information will be browsed by the domain experts*, and so on. In the third part, we discussed with the domain expert the problems in the current approach, the existing challenges, their requirements on a new approach for model selection, and the opportunities that visual analytics techniques can bring about.

During the interview, the domain experts complained about the difficulty in model selection due to the large number of data they needed to explore. The current approach presents the forecast accuracy of each model for all the products in a tabular form, which is often overwhelming and requires a long scroll to explore the data. Also, the domain experts commented that they were not sure whether the selected model could produce an accurate forecast or not. In addition, they agreed with us on supporting model selection based on the products with similar historical demand. They commented that it follows the rule of product lifecycle [50], and hoped that the new approach can make full use of previous forecasting experience.

Design Requirements
Based on the feedback collected from the interview study, we summarized the following five design requirements.

R1. Allow interactive model ranking. The system should support the interactions that allow users to adjust the weights of different features for ranking demand forecasting models. In addition, it should also enable users to manually remove dissimilar products identified based on the features of the historical demand data extracted by automated algorithms.

R2. Enable model selection based on the forecast for similar products. The system should allow users to interactively select the products with similar historical demand and help them find the models with the best performance on these products. This can take advantage of previous forecasting experience and alleviate the problem that a model selected according to the average performance on all the products may produce an inaccurate forecast for some products.

R3. Present the performance of models on different products. Besides the overall performance of models on the selected products, the system should also show the performance of models on individual products, which is helpful for users to identify the drawback of a model and seek alternative models.

R4. Reveal the risk of an inaccurate forecast. When applying a demand forecasting model, users have no idea whether the forecast is reliable or not. Revealing the risk of an inaccurate forecast can prevent the adoption of an unreliable model and increase the confidence in model selection.
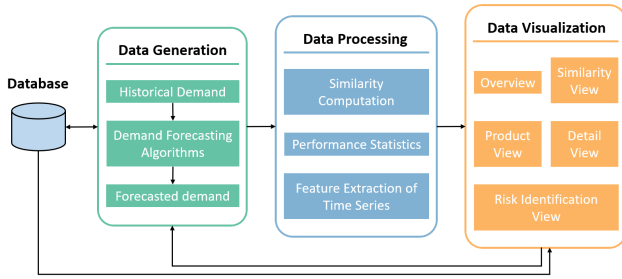
Figure 1. An overview of the DFSeer system, which mainly consists of three modules: data generation, data processing and data visualization.

R5. Support multi-level exploration. Displaying the large quantity of data related to demand forecasting on a limited screen is overwhelming and can result in visual clutter. The system should support the exploration of demand forecasting data with levels of details to reduce the manual effort required.

DFSEER

In this section, we introduce DFSeer, an interactive visual analytics system that supports model selection for demand forecasting and reveals the risk of the selected model. This section first gives an overview of DFSeer and then describes each component of the visualization system in detail.

System Overview

Figure 1 presents an overview of the DFSeer system, which can be divided into three modules: data generation, data processing, and data visualization.

The data generation module forecasts the demand for products in the next month with various kinds of forecasting algorithms. Three analytical tasks are then performed in the data processing module. First, the Euclidean Distance [21] is utilized to measure the similarity between the z-normalized [27] historical demand data. Second, we compute the performance indicators to evaluate and rank different demand forecasting models. These performance indicators include the accuracy of the forecast in the next month, the variance of the forecast accuracy in the recent ten months, and the number of products to which a demand forecasting model is applicable. To be more specific, a model is applicable to a product when it is among the most accurate k models, where k is defined by users. Third, automated algorithms are introduced to extract the properties of the historical demand data, including the trend, the seasonality, the autocorrelation, and the stationarity. The demand data and the information generated by the data processing module are then sent to the visualization module.

There are five components in the data visualization module: (a) an overview which allows rerunning the algorithm to forecast the demand in another month, supports interactively adjusting the weight for ranking demand forecasting models, and summarizes the performance of different models on the products selected by users, (b) a similarity view which encodes the similarity between time series with the distance between the points in a scatter plot and allows users to select clusters for detailed analysis, (c) a product view which displays the

detailed performance of models on individual products with a novel glyph design, (d) a detail view which presents the historical demand data of a user-specified product and the accuracy of the forecasts in past months by the top k models, and (e) a risk identification view which shows the features of similar historical demand data, and assists users in evaluating the risk of a model. The visualization module also supports several interactions such as linked-filtering and the tooltip.

With DFSeer, the user can compare and select different types of demand forecasting models with levels of details. Besides, the methodology applied in model selection (i.e., model selection is based on the performance of models on the products with similar historical demand data) can increase the confidence of users in the selected model.

Ranking Models Interactively

DFSeer enables users to adjust the weight of the performance indicators for ranking demand forecasting models (**R1**). The models are re-ranked according to the adjustment, with the summarized performance indicators of each model presented in a ranked list. This presents the macroscopic information to support model selection for demand forecasting (**R2**, **R5**).

As shown in the upper part of the overview (Figure $2a_1$), three sliders are provided for users to adjust the weight of the forecast accuracy, the variance of the forecast accuracy in recent months, and the number of products to which a demand forecasting model is applicable. There is another slider for parameter "top k", which indicates that among the models forecasting the demand of a product, only the most accurate k models are considered applicable to that product. The ranked list of the demand forecasting models is shown below the sliders (Figure $2a_2$), which summarizes the model performance on the products selected by users. For each model, the leftmost and rightmost rightward bars display the accuracy of the forecast and the number of the products to which a demand forecasting model is applicable respectively. The circle in the middle encodes the variance of the forecast accuracy in recent months with the color saturation, which is consistent with other views. Only the top k models in the ranked list are depicted by different colors other than gray.

Selecting Similar Products

The similarity view (Figure $2b$) allows users to select the products with similar historical demand data in a scatter plot (**R2**). We employ the Euclidean Distance [21] to measure the similarity between the z-normalized [27] time series. The Euclidean Distance is adopted because it is efficient [18] and can handle noisy data. Then, multidimensional scaling (MDS) [7] is applied to map the time series to a scatter plot, where the relative distance between two points indicates the similarity between two time series. In the scatter plot, the red points represent the time series to be forecasted, while the light purple points encode the time series forecasted before. Users can select a group of adjacent points by drawing a lasso to explore the performance of models on individual products.

Exploring Model Performance on Products

After selecting a cluster of time series of interest, users can view the performance of models on individual products and
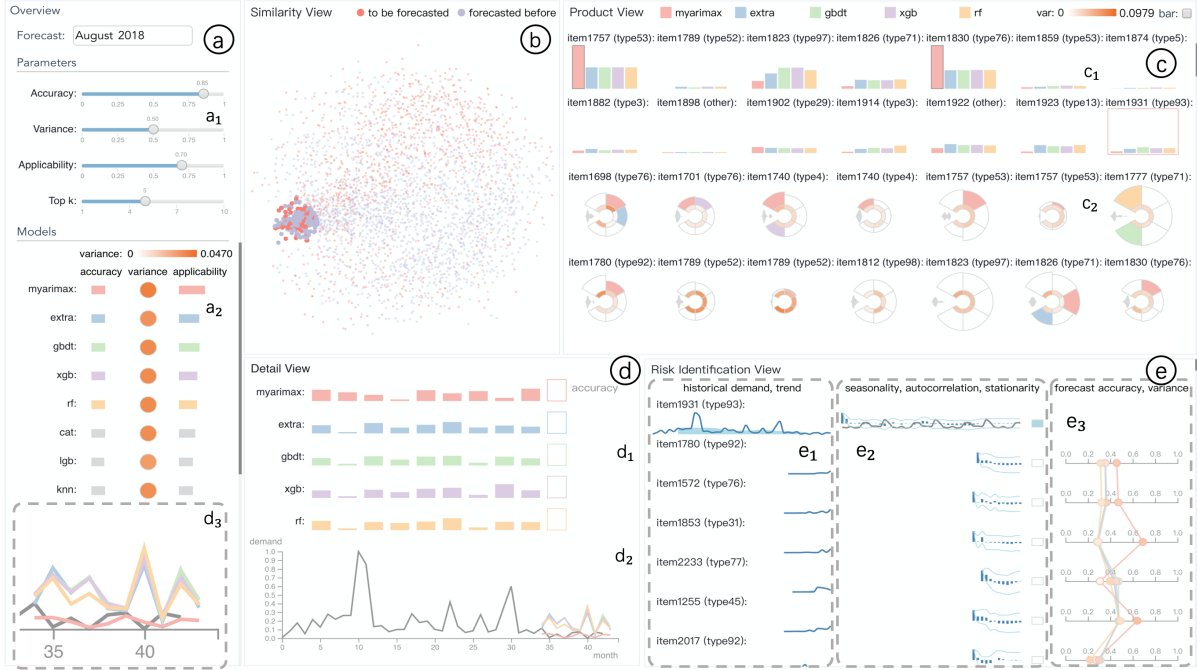
Figure 2. DFSeer supports interactive model selection for demand forecasting and reveals the risk of the selected model. **(a)** The overview enables users to adjust the parameters for ranking demand forecasting models (a_1) and presents the ranked list of models (a_2). **(b)** The similarity view allows users to select the products with similar historical demand data in a scatter plot. **(c)** The product view encodes the model performance on individual products with bar charts (c_1) and carefully-designed glyphs (c_2). **(d)** The detail view shows the monthly forecast accuracy of the top k models (d_1), and the real and forecasted demand of a user-selected product (d_2). The right part of d_2 is enlarged and shown in d_3. **(e)** The risk identification view presents the extracted properties of the historical demand (e_1, e_2) and the model performance difference on similar products to reveal the risk of the selected model.

further explore the forecast details of a user-specified product in recent months (**R3**). This enables users to select demand forecasting models by comparing their mesoscopic and microscopic information (**R2**, **R5**).

For the time series to be forecasted, the product view (Figure 2c) employs a bar chart (Figure 2c_1) to display the forecasted value of the top k models in the overview. The order of the bars is the same as that in the ranked list. For the time series forecasted before, we exploit a well-crafted glyph design (Figure 2c_2) to describe the performance of the forecasts by different models. The current visual design is inspired by prior studies on multi-dimensional data visualization [62, 61, 12, 57] and we worked closely with domain experts to collect their feedback on the visual design in an iterative manner. The glyph takes a circular design, which encodes the most accurate k models by different sectors. The sectors are arranged clockwise according to the forecast accuracy. In each sector, the inner arc depicts the variance of the forecast accuracy in recent months with the color saturation, while the outer arc represents the accuracy of the forecast by the radius. The color of the outer arc encodes the model type. In the left part of the glyph, a violin plot shows the accuracy distribution of other models. However, the circular glyph design is not very convenient for comparing model performance on different products. DFSeer also supports a bar chart-based glyph design (Figure 3b), where the upper bars depict the forecast accuracy of the top k models and the lower rectangles encode the variance of the forecast accuracy with the color saturation. A marginal density plot on the right shows the accuracy distribu-
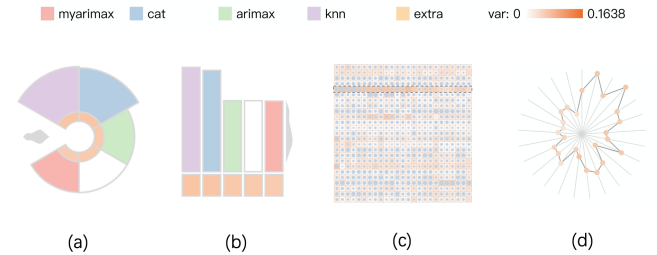

Figure 3. The glyph designs to describe the model performance on each product: **(a)** the circular glyph design used in DFSeer, **(b)** a bar chart-based glyph design, **(c)** a matrix view-based glyph design, and **(d)** a star plot-based glyph design.

tion of other models. Since the circular glyph design can help users distinguish the performance indicators on the products forecasted before (Figure 2c_2) from the forecasting results of future product demand (the bar chart shown in Figure 2c_1), and it is also visually attractive, DFSeer uses the circular glyph design as the default choice. But users can interactively switch to the bar chart-based glyph design.

Two alternatives of the glyph are considered in our iterative design process, as shown in Figure 3. Our earliest attempt is to use a matrix view (Figure 3c) to describe the performance of models on individual products. Each row of the matrix represents a product, while each column represents a model. However, much manual effort is required to find the most suitable models for each product. Next, a star plot (Figure 3d) is adopted to show the performance of different models, where

each star denotes a product, and each spoke denotes the performance of a model. However, showing dozens of spokes is overwhelming and it is also not easy to identify the most accurate k models for each product in the star plot.

After users click a bar chart (i.e., the time series to be forecasted) in the product view, the detailed performance of the top k models, the actual demand data and the latest forecast in the past months will be shown in the detail view (Figure 2d). Similar to the overview, a bar chart (Figure 2d_1) is used to express the accuracy of the forecasts in the past months. The horizontal axis of the bar chart encodes the month, while the height encodes the accuracy. The bar charts are vertically aligned, whose order and color are consistent with those in the overview. In addition, users can also inspect the raw data (i.e., the actual and forecasted demand) in a multi-line chart below the bar charts (Figure 2d_2).

Identifying the Risk of the Selected Model
When users select a model to forecast the demand, it is critical to reveal the risk of the selected model. The risk identification view (Figure 2e) is proposed to extract the properties of the time series (i.e., the historical demand of products) that are similar to the one that users specify in the product view and present the performance differences of the top k models on these time series. It can help users identify the probability that the selected model will produce an accurate forecast (**R4**).

To find the properties of a time series that are important for the forecast, we study how classical time series models make a forecast [34, 17] and discuss with domain experts how they judge whether two time series are similar. We then identify four key properties of a time series, i.e., the trend, the seasonality, the autocorrelation, and the stationarity. To extract the properties of a time series, we first apply an additive model [14] to decompose a time series to get the trend and seasonality components. Later, the Autocorrelation Function (ACF) plot [9] is utilized to analyze the correlation between the demand in different months. Additionally, the augmented Dickey-Fuller (ADF) test [13] can reveal the stationarity of a time series.

In the risk identification view, the first row presents the properties of the time series to be forecasted, which is specified by users. The following rows show the properties of the time series forecasted before and the performance of the top k models, which are ranked by the similarity between the time series and the user-specified one. In the left part of the risk identification view (Figure 2e_1), a line chart shows the historical demand, overlaid with an area chart to display the trend of the time series. Figure 2e_2 utilizes a black line to exhibit the seasonality of a time series, which is overlaid with a bar chart to present the autocorrelation function. The two light blue lines indicate the 95% confidence interval of the autocorrelation. Besides, a rectangle is placed in the right part of Figure 2e_2 to encode the p-value output by the ADF test. The rectangle is green if the historical demand data is stationary, or white if the data is not stationary. A vertically arranged parallel coordinates plot in the right part of the risk identification view (Figure 2e_3) shows the performance of the top k models in forecasting the demand of the similar products. In the parallel coordinates plot, each horizontal axis represents a time series forecasted
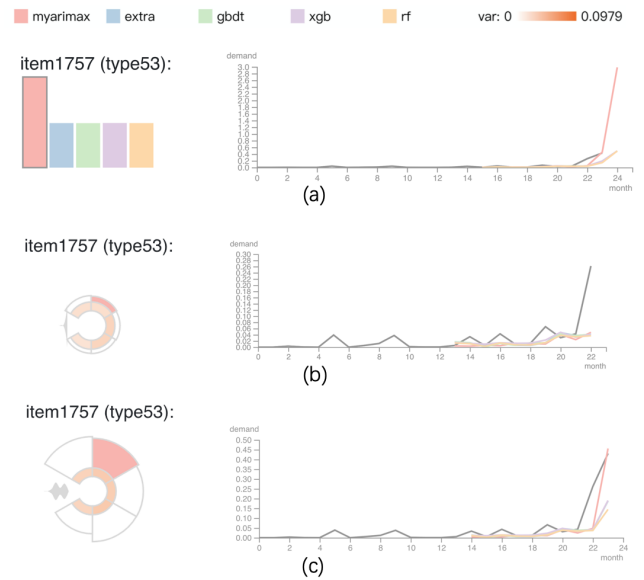

Figure 4. (a) *item1757* gains a significantly high demand forecasting value by *myarimax* in August 2018. (b) An abrupt sharp increase in demand in June 2018 results in low forecast accuracy. (c) *myarimax* achieves a much better demand forecasting accuracy in July 2018 than other top models, indicating its excellent sensitivity in learning an abrupt demand increase.

before, while each of the top k models is denoted as a polyline. The intersection of the polyline and the axis is displayed as a circle, whose horizontal position encodes the forecast accuracy and color saturation encodes the variance of the forecast accuracy in recent months. We encode the variance with the color saturation instead of the size of a circle or a bar to avoid possible occlusion. When exploring the time series forecasted before and the performance of the top k models, users are allowed to remove a time series if they think it is dissimilar to the time series to be forecasted.

CASE STUDY
In this section, we conducted two case studies to demonstrate the effectiveness of DFSeer. The users involved in the case studies are the domain experts from the manufacturing company we collaborate with, who also attended the interview study in the next section. The dataset used consists of monthly demand of 1,096 products from January 2015 to November 2018. The actual names and types of the products are removed due to privacy issues.

Inspecting Top Models: Best Model Does Not Work for All
In the first case study, a user was asked to select a suitable model to forecast the demand in August 2018 for any cluster of products of his interest. After loading the data to DFSeer and setting the forecast period, he immediately found a cluster from the similarity view (Figure 2b). The user first set the parameter *top k* to 5 in the overview and filtered the data to gain a rough idea of the model performance (**R5**). From the ranked list of models, the user quickly noticed that the forecast accuracy of different models is a bit low. Therefore, he further increased the weights of accuracy and applicability to check which model can lead to higher average accuracy (**R1**).

Figure 5. Identify the difference between "similar" models by using DFSeer: The user found that the top 2 models (i.e., *cat* and *knn*) have very similar average performance (a_2) and an overall similar distribution of ranking among the top 5 models for the selected products (c_2). He then clicked an example product of *type76* (i.e., *item1612*) and found the two models have similar variance of the monthly accuracy (d_1) and a close trend of the forecast accuracy for the similar products (*e*). But the forecasted demand by *cat* is consistently lower than that by *knn* (d_2 or d_3). To mitigate the shortage of this type of key products, he selected *knn* to forecast the demand for the products of *type76*.

The ranked list (Figure $2a_2$) shows that the model *myarimax* (e.g., a manually tuned ARIMAX model [60]) achieves the highest average accuracy. Meanwhile, the circular glyphs in the product view (Figure $2c_2$) also indicate that *myarimax* consistently ranks as one of the top 5 models when forecasting the demand of the majority of the similar products shown in the product view (**R3**). By quickly clicking the similar products to explore the performance details of the top 5 models, the user observed that *myarimax* is more accurate than other top models when forecasting the demand in different months (Figure $2d_1$) and the forecasted demand by the model is closer to the real one (Figure $2d_2$ or $2d_3$). He also compared the performance of the top 5 models on similar products in the risk identification view and found that *myarimax* is more accurate than other top models for most products (Figure $2e_3$). Thus, the user became more confidence in *myarimax* and decided to consider *myarimax* as a candidate model for forecasting the demand of most of the selected products (**R2**).

However, the user also quickly identified some abnormal forecasts from the product view. The red bars of two products (i.e., *item1757* and *item1830*) are surrounded by a black border, as shown in Figure $2c_1$. It means that the product demand forecasted by *myarimax* exceeds a threshold predefined based on domain knowledge. So the user chose one of the two products (i.e., *item1757*) for further analysis. By looking at the detail view, the user found that *item1757* and other products of the same type (i.e., *type53*) usually have small demand values, but only *item1757* gains a significantly high demand forecasting value by *myarimax* (Figure 4a). Such an abnormal observa-

tion gives the user the impression that the forecasting result of *item1757* by *myarimax* may be problematic. By further checking the historical demand of *item1757*, the user easily observed that *item1757* has an abrupt sharp increase in demand in June (Figure 4(b)) and July (Figure 4(c)). *myarimax* achieves a much better demand forecasting accuracy in July than other top models, indicating its excellent sensitivity in learning an abrupt demand increase. However, it overrates this trend and forecasts an abnormally high product demand of *item1757* in August. For the other product (i.e., *item1830*), the user also observed similar issues.

These observations from DFSeer informed the user of the potential risk of using *myarimax* to forecast product demand (**R4**). Although it is the "best" model in terms of the overall performance, it can also overrate a sudden increase in product demand, resulting in demand forecasting failures.

Breaking the Tie: "Similar" Models Can Be Different
In this case study, another user was asked to finish exactly the same task as the first case study. But he selected a different group of similar products by brushing in the similarity view, as is shown in Figure 5b.

By browsing the product view (Figure $5c_1$ and $5c_2$), the user easily noticed that many products are belonging to *type76*. Since the products of *type76* are the sub-components for producing other products and play a critical role in the whole supply chain of the factory, a demand forecasting model that is both accurate and stable will be preferred. Therefore, the

user significantly increased the weight of accuracy and variance (Figure $5a_1$) to favor the demand forecasting models that have good average forecast accuracy and also perform consistently well along the time (**R1**). The ranked list of models (Figure $5a_2$) shows that the model *cat* (i.e., CatBoost [19]) and the model *knn* (i.e., the KNN model [3]) rank the first and second, respectively. But the two models have very **similar** performance in terms of the average accuracy and variance, making it difficult to distinguish which model is actually more appropriate for the selected products. Therefore, he further checked the performance distribution of the top five models on the selected products whose demand was forecasted before (**R3**, **R5**). But the user still found that an overall **similar** distribution of *cat* and *knn* models that rank within the top five models for the selected products, as shown in Figure $5c_2$.

However, when the user continued to closely explore the product view, he quickly observed that *cat* (the red bars) tends to always forecast lower future demand for the selected products (especially the products of *type76*) than *knn* (the blue bars), as shown in Figure $5c_1$. This phenomenon is beyond the expectation of the user, since both models achieved **similar** accuracy and variance for different products. Therefore, he selected a product of *type76* (i.e., *item1612*) from the product view to inspect more details. The detailed view shows that *cat* and *knn* have a very similar demand forecasting accuracy along the time for *item1612* (Figure $5d_1$), and their demand forecasting results for other products with similar historical demand share a close trend (Figure 5e). However, when the user carefully checked the line charts of the forecasting results in the detail view (Figure $5d_2$ or $5d_3$), he clearly observed that the demand forecasting result by *cat* is **consistently lower** than that by *knn* along the time regardless of the drop and rise of the actual demand (**R4**). Such an observation was also found for other products of *type76*.

As mentioned above, the products of *type76* are critical in the whole supply chain of the factory. The shortage of these products can bring serious trouble to the manufacturing of the factory. Therefore, despite the similarity between *cat* and *knn*, the user finally decided to choose *knn* as the demand forecasting model for the products of *type76*, since it often has higher demand forecasts than *cat* (**R2**). Though such a choice can increase the risk of overstock, it can better mitigate the shortage of these key products.

EXPERT INTERVIEW
We conducted semi-structured interviews with four domain experts working on model selection for demand forecasting to demonstrate the effectiveness and usability of DFSeer. This section first introduces the interview procedures. We then summarize the feedback collected from the experts.

Participants and apparatus
The participants include four domain experts from the manufacturing company we work with. The domain experts are different from the experts attending the requirement analysis interview and haven't participated in the design of the DFSeer system. Two of the domain experts (E_1 and E_2) are algorithm

Q1	Find the most accurate demand forecasting model.
Q2	Find the most stable demand forecasting model.
Q3	Find the demand forecasting model which has the best applicability.
Q4	Find the model whose forecast for the target month is different from other top k models.
Q5	Find the product for whom the forecast by the top k models is the most accurate.
Q6	For a given product, find the most unstable model among the top k models.
Q7	For a given product, find the model whose forecast is the closest to the real demand.
Q8	Find the product whose historical demand is the most dissimilar to the selected one.
Q9	Given the products which are similar to a selected one, find the product whose forecasted demand is the most accurate.

Table 1. The questions in T1 are related to model performance in different levels of details: the overview (Q1 - Q3), the product view (Q4 - Q5), the detail view (Q6 - Q7), and the risk identification view (Q8 - Q9).

developers who are working on developing demand forecasting algorithms, while the other two (E_3 and E_4) are demand analysts who are responsible for selecting the most suitable demand forecasting models for different products from dozens of algorithms provided by algorithm developers. The interviews were conducted with a 23.8-inch 1920 x 1080 monitor showing the visual analytics system.

Dataset and Tasks
The dataset used in the expert interviews is the same as that in the case studies, which contains the monthly demand data of 1,096 products from January 2015 to November 2018. During the interviews, we asked the domain experts to fulfill two carefully-designed tasks to assess the usability of the visualization system and understand how the domain expert selects a demand forecasting model.

- **T1.** Analyze the result of model selection for a group of similar products.
- **T2.** Select suitable models to forecast the demand of pre-selected products.

In **T1**, we plan to select a model to forecast the demand of products in August 2018 and pre-select a group of similar products. The corresponding model performance on these products is displayed in the visualization system. The domain experts are then asked to answer nine questions related to the model performance in different levels of details, as shown in Table 1. In **T2**, the domain experts are required to find the best model for two types of products, i.e., the general products which emphasize more on the accuracy of a demand forecasting model (e.g., switches and indicator lights), and the products that are important to the supply chain of the factory and need a more stable model (e.g., high-speed optical modules and digital signal controllers).

Procedure
During the interview, we first introduced the methodology of demand forecasting model selection, the visual design and

Q1	The model selection based on the products with similar historical demand is reasonable.
Q2	The workflow provides useful guidance for me to select suitable demand forecasting models.
Q3	The workflow can reduce my workload in model selection for product demand forecasting.
Q4	I have confidence in the model selected by using the visualization system.
Q5	The visual design is intuitive.
Q6	The visual design is easy to understand.
Q7	The visual design provides enough information for me to select a demand forecasting model.
Q8	The visual design and interactions help me compare and select demand forecasting models.
Q9	The visual analytics system is easy to use.
Q10	The visual analytics system is helpful for demand forecasting model selection.
Q11	I would like to use the visual analytics system to select demand forecasting models in the future.
Q12	I will recommend the visual analytics system to my colleagues working on forecasting model selection.

Table 2. The questionnaire consists of three parts: the workflow and trust (Q1 - Q4), the visual design and interactions (Q5 - Q8), and the usability (Q9 - Q12).

interactions, and the functionality of each part of the DFSeer system to the domain experts. After that, an example usage scenario was given to help them understand how to use the visualization system to select demand forecasting models. Then, the participants were allowed to explore the system on their own. The aforementioned tutorial lasted for about 20 minutes. After the tutorial, the domain experts were invited to discuss how they select a demand forecasting model and what information is used to support their decision in their daily work. Then, the participants were guided to carry out two tasks. Both tasks employed a think-aloud protocol, and the comments and suggestions of the domain experts were all recorded. After accomplishing the two tasks, the participants were asked to describe the workflow of using the DFSeer system to select demand forecasting models. They were also encouraged to compare DFSeer with the approach currently used in their daily work. Then, we invited the participants to rate the visual analytics system on a 7-point Likert scale (1 - strongly disagree, 7 - strongly agree) from three aspects. The detailed questions are shown in Table 2. Finally, the domain experts attended a post-study interview where they gave comments and suggestions on the workflow and trust, the visual design and interactions, and the usability of the DFSeer system. The study for each domain expert lasted for approximately one hour and 20 minutes.

Results

We summarize the feedback of the interviews and the result of the questionnaire into three parts:

Workflow and trust. All the domain experts agreed that the workflow for demand forecasting model selection in our work is helpful and increases their trust in the selected model. E_3 said "In our daily work, model selection is based on the

average accuracy of a model. It is difficult for me to trust that the model would give an accurate forecast for a certain product." He commented the proposed workflow to support the selection of models based on their performance on similar products is reasonable and trustful, since the algorithm will often produce similar results for similar time series data [10, 4, 21] and the change of demand over months follows the rule of product lifecycle [50]. E_2 mentioned "Having seen the detailed model performance on similar products, I have a better understanding of the models and trust the model I selected." E_4 reported it was difficult to manually identify the products with dissimilar historical demand. With the properties extracted and presented in the risk identification view, he was able to find dissimilar time series without much effort. When performing **T2**, E_1 found the model he selected is different from that by others. After checking the risk identification view, he realized many dissimilar products were included. He then shrunk the scope of selecting similar products and got a suitable model. E_1 commented "The risk identification view helps me check whether the scope of selecting similar products is suitable."

The domain experts also gave suggestions on the selection of products with similar historical demand data. E_2 suggested involving the relationship between different types of products in the computation of similarity. E_4 mentioned there are too many points in the similarity view and recommended that the system can further filter products by the type of products and the relationship between them.

Visual design and interactions. The domain experts reported that the visual design is informative and the interactions are user-friendly. E_3 compared DFSeer with the approach currently employed in their work and commented "DFSeer provides an intuitive and easy way for me to explore the performance of models. Without it, I need to scroll a large tabular form to search for the information I need." E_4 liked the glyph design in the product view and mentioned "The bar chart can reveal which model produces a different forecast, while the circular glyph shows which model is more accurate or stable." He also stated that the glyph design is in line with their daily work, where the domain experts only focused on several top-ranked models. The domain experts were surprised by the capability of the risk identification view. E_1 found there were several products with few historical demand records and removed them from the risk identification view. He commented "the tool combines the power of automated algorithms and our knowledge to judge whether a product is similar or not". E_2 mentioned the interactions are easy and powerful. He said "Through the interactions with the visualization system, including adjusting the weights for ranking models, selecting similar products, and removing dissimilar historical demand data, I get highly involved in how the models are ranked and gain a better understanding of the models."

Despite the positive feedback, the domain experts commented that the product view and the risk identification view had a steep learning curve, especially for the users who have little knowledge of visualization.

Usability. The domain experts confirmed that the DFSeer system is easy to use and can facilitate their work in model se-

lection for demand forecasting. It took them about 25 minutes to finish the tasks (including the task introduction), where no task was skipped. Only one participant made an error when answering the objective questions in **T1** (i.e., Q5), which is because there are two products with very similar model performance. This shows that DFSeer helps analyze the model performance on similar products. E_4 said "Compared with our daily work where the similarity between products is manually determined based on the type of products, DFSeer provides a more reasonable and flexible manner for finding similar products." The domain experts agreed that DFSeer provides helpful guidance for them to find an appropriate demand forecasting model and lightens their workload a lot. E_3 mentioned the visual analytics system was easy to learn and use. He appreciated the nice experience of using the system, which was better than using the tedious tabular form. The domain experts planned to deploy the DFSeer system in the manufacturing factory to facilitate their daily work of model selection. They agreed it is worth spending around 20 minutes learning to use the system.

Besides the positive comments on the usability of the DFSeer system, the domain experts also put forward several constructive suggestions. E_4 expressed that the system could involve other advanced similarity calculation methods and allow users to choose which method to use. E_1 suggested there was a need for post-processing which can provide a feasible demand value when the selected model offers an abnormal forecast.

DISCUSSION
Although DFSeer has shown its ability in facilitating model selection for demand forecasting, limitations still exist. In this section, we discuss the limitations from the perspectives of visual design scalability, lessons learned, and generalizability and extensions.

Visual Design Scalability
The similarity view of the DFSeer system displays 3,288 time series for users to select similar products. To present a larger number of time series, measures shall be taken to reduce the overlap of data points. For example, we can adopt a hierarchical strategy which first asks users to assign the type of products and specify the range of distance for searching similar products, and then shows the products inside this range. The visual analytics system only presents the detailed performance of less than ten models. This is in accordance with the workflow of selecting demand forecasting models in the manufacturing factory, where domain experts only focus on several models with the best performance. Showing the detailed performance of all the models could increase the visual burden of users.

Lessons Learned
By working with domain experts, we realize that showing the internal structure of models to demand analysts may not be that useful. In our study, the demand analysts often do not have a strong background in machine learning models. Hence, instead of showing the internal structure of models, DFSeer visualizes the patterns of the input (i.e., time series), the output (i.e., forecasted demand and performance indicators), and the correlations between them. Besides, we find the model

ranked first by the algorithm may not always be the best. For example, in our second case study, the domain experts selected the model ranked second because compared with the model ranked first, it can relieve the shortage of the products that are critical to the supply chain. In addition, our work shows that model transparency is highly appreciated and needed in model selection for demand forecasting. Instead of only being told to use a model, users also want to know why some time series are considered similar and how the model performance is on these similar products.

Generalizability and Extensions
Although in our study, DFSeer is used to support model selection for demand forecasting, it can be adapted to other application scenarios, such as stock market prediction, power usage prediction, traffic flow forecast, and medical diagnosis. For example, in the scenario of power usage prediction, DFSeer can group residents based on their residence locations and incomes, and help power managers find suitable prediction models for different groups of residents and identify potential power shortages. In the scenario of medical diagnosis, DFSeer can be adapted to diagnosing the illness according to the cases with similar symptoms, which can help doctors make a diagnosis by summarizing the predictions of multiple models. DFSeer can also be extended to increase its power in model selection for demand forecasting. First, other competitive time series forecasting models can be involved in our system. Second, the relationship between different types of products can be considered when computing the similarity between time series. Third, the system can also provide more distance calculation strategies for users to choose from.

CONCLUSION AND FUTURE WORK
We have presented DFSeer, a visual analytics system to help data analysts select demand forecasting models based on the products with similar historical demand. The visualization system provides different levels of details for users to explore and compare the performance of demand forecasting models. Besides, the DFSeer system can reveal the risk of a model by showing the performance difference of the model on similar products. We evaluated the visual analytics system through two case studies with product demand data and semi-structured interviews with four domain experts working on model selection for demand forecasting in a manufacturing factory, which shows that DFSeer is helpful and effective in facilitating model selection for demand forecasting.

In future work, we plan to include more time series forecasting algorithms in the visualization system. Besides, we would like to consider the relationship between different types of products when computing the distance between the historical demand data. We also want to enable users to filter products by type in the similarity view. Finally, we intend to provide other advanced approaches to compute the distance between time series and allow users to select the approach on their own.

REFERENCES

[1] Nesreen K Ahmed, Amir F Atiya, Neamat El Gayar, and Hisham El-Shishiny. 2010. An empirical comparison of machine learning models for time series forecasting. *Econometric Reviews* 29, 5-6 (2010), 594–621.

[2] Wolfgang Aigner, Silvia Miksch, Heidrun Schumann, and Christian Tominski. 2011. *Visualization of time-oriented data*. Springer Science & Business Media.

[3] Naomi S Altman. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician* 46, 3 (1992), 175–185.

[4] Francisco Martinez Alvarez, Alicia Troncoso, Jose C Riquelme, and Jesus S Aguilar Ruiz. 2010. Energy time series forecasting based on pattern sequence similarity. *IEEE Transactions on Knowledge and Data Engineering* 23, 8 (2010), 1230–1243.

[5] Benjamin Bach, Pierre Dragicevic, Daniel Archambault, Christophe Hurter, and Sheelagh Carpendale. 2014. A review of temporal data visualizations based on space-time cube operations.

[6] Markus Bögl, Wolfgang Aigner, Peter Filzmoser, Tim Lammarsch, Silvia Miksch, and Alexander Rind. 2013. Visual analytics for model selection in time series analysis. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2237–2246.

[7] Ingwer Borg and Patrick Groenen. 2003. Modern multidimensional scaling: Theory and applications. *Journal of Educational Measurement* 40, 3 (2003), 277–280.

[8] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. 2015. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons.

[9] Peter J Brockwell and Richard A Davis. 2016. *Introduction to Time Series and Forecasting*. springer.

[10] Paolo Buono, Catherine Plaisant, Adalberto Simeone, Aleks Aris, Ben Shneiderman, Galit Shmueli, and Wolfgang Jank. 2007. Similarity-based forecasting with simultaneous previews: A river plot interface for time series forecasting. In *2007 11th International Conference Information Visualization (IV'07)*. IEEE, 191–196.

[11] Stef van Buuren and Miranda Fredriks. 2001. Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in Medicine* 20, 8 (2001), 1259–1277.

[12] Yuanzhe Chen, Qing Chen, Mingqian Zhao, Sebastien Boyer, Kalyan Veeramachaneni, and Huamin Qu. 2016. DropoutSeer: Visualizing learning patterns in Massive Open Online Courses for dropout reasoning and prediction. In *2016 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 111–120.

[13] Yin-Wong Cheung and Kon S Lai. 1995. Lag order and critical values of the augmented Dickey–Fuller test. *Journal of Business & Economic Statistics* 13, 3 (1995), 277–280.

[14] William P Cleveland and George C Tiao. 1976. Decomposition of seasonal time series: A model for the Census X-11 program. *J. Amer. Statist. Assoc.* 71, 355 (1976), 581–587.

[15] Timothy J Cole and Pamela J Green. 1992. Smoothing reference centile curves: the LMS method and penalized likelihood. *Statistics in Medicine* 11, 10 (1992), 1305–1319.

[16] R Dennis Cook and Sanford Weisberg. 1982. *Residuals and influence in regression*. New York: Chapman and Hall.

[17] Alysha M De Livera, Rob J Hyndman, and Ralph D Snyder. 2011. Forecasting time series with complex seasonal patterns using exponential smoothing. *J. Amer. Statist. Assoc.* 106, 496 (2011), 1513–1527.

[18] Hui Ding, Goce Trajcevski, Peter Scheuermann, Xiaoyue Wang, and Eamonn Keogh. 2008. Querying and mining of time series data: experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment* 1, 2 (2008), 1542–1552.

[19] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. 2018. CatBoost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363* (2018).

[20] Christos Faloutsos, Jan Gasthaus, Tim Januschowski, and Yuyang Wang. 2018. Forecasting big time series: old and new. *Proceedings of the VLDB Endowment* 11, 12 (2018), 2102–2105.

[21] Christos Faloutsos, Mudumbai Ranganathan, and Yannis Manolopoulos. 1994. *Fast Subsequence Matching in Time-series Databases*. Vol. 23. ACM.

[22] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. 2014. Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research* 15, 1 (2014), 3133–3181.

[23] Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, and Frank Hutter. 2015. Efficient and robust automated machine learning. In *Advances in Neural Information Processing Systems*. 2962–2970.

[24] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 1126–1135.

[25] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics* (2001), 1189–1232.

[26] Jonah Gabry, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. 2019. Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 182, 2 (2019), 389–402.

[27] Dina Q Goldin and Paris C Kanellakis. 1995. On similarity queries for time-series data: constraint specification and implementation. In *International Conference on Principles and Practice of Constraint Programming*. Springer, 137–153.

[28] Steve R Gunn and others. 1998. Support vector machines for classification and regression. *ISIS Technical Report* 14, 1 (1998), 5–16.

[29] Anshuman Gupta and Costas D Maranas. 2003. Managing demand uncertainty in supply chain planning. *Computers & Chemical Engineering* 27, 8-9 (2003), 1219–1227.

[30] H2O.ai. 2017. AutoML. `http://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html`. (2017). Accessed: 2019-08-11.

[31] Arthur E Hoerl and Robert W Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12, 1 (1970), 55–67.

[32] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M Drucker. 2019. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 579.

[33] Rob Hyndman, Anne B Koehler, J Keith Ord, and Ralph D Snyder. 2008. *Forecasting with Exponential Smoothing: the State Space Approach*. Springer Science & Business Media.

[34] Rob J Hyndman, Yeasmin Khandakar, and others. 2007. *Automatic Time Series for Forecasting: the Forecast Package for R*. Number 6/07. Monash University, Department of Econometrics and Business Statistics âĂe.

[35] Lars Kotthoff, Chris Thornton, Holger H Hoos, Frank Hutter, and Kevin Leyton-Brown. 2017. Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. *The Journal of Machine Learning Research* 18, 1 (2017), 826–830.

[36] Josua Krause, Adam Perer, and Enrico Bertini. 2014. INFUSE: interactive feature selection for predictive modeling of high dimensional data. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 1614–1623.

[37] Josua Krause, Adam Perer, and Kenney Ng. 2016. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5686–5697.

[38] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436.

[39] Hau L Lee. 2002. Aligning supply chain strategies with product uncertainties. *California Management Review* 44, 3 (2002), 105–119.

[40] Lezhi Li, Yunfeng Bai, and Yang Wang. 2019. Manifold: A Model-Agnostic Visual Debugging Tool for Machine Learning at Uber. (2019).

[41] Mengchen Liu, Jiaxin Shi, Zhen Li, Chongxuan Li, Jun Zhu, and Shixia Liu. 2016. Towards better analysis of deep convolutional neural networks. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2016), 91–100.

[42] Yafeng Lu, Robert Krüger, Dennis Thom, Feng Wang, Steffen Koch, Thomas Ertl, and Ross Maciejewski. 2014. Integrating predictive analytics and social media. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 193–202.

[43] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*.

[44] Yao Ming, Shaozu Cao, Ruixiang Zhang, Zhen Li, Yuanzhe Chen, Yangqiu Song, and Huamin Qu. 2017. Understanding hidden memories of recurrent neural networks. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 13–24.

[45] mljar. 2018. mljar. `https://mljar.com/`. (2018). Accessed: 2019-08-11.

[46] Wolfgang Müller and Heidrun Schumann. 2003. Visualization for modeling and simulation: visualization methods for time-dependent data-an overview. In *Proceedings of the 35th Conference on Winter Simulation: Driving Innovation*. Winter Simulation Conference, 737–745.

[47] William Playfair. 1801. *The Commercial and Political Atlas: Representing, by Means of Stained Copper-plate Charts, the Progress of the Commerce, Revenues, Expenditure and Debts of England during the whole of the Eighteenth Century*. T. Burton.

[48] Takafumi Saito, Hiroko Nakamura Miyamura, Mitsuyoshi Yamamoto, Hiroki Saito, Yuka Hoshiya, and Takumi Kaseda. 2005. Two-tone pseudo coloring: Compact visualization for one-dimensional data. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005*. IEEE, 173–180.

[49] Xingjian Shi and Dit-Yan Yeung. 2018. Machine learning for spatiotemporal sequence forecasting: A survey. *arXiv preprint arXiv:1808.06865* (2018).

[50] John Stark. 2015. Product lifecycle management. In *Product Lifecycle Management (Volume 1)*. Springer, 1–29.

[51] Dong Sun, Renfei Huang, Yuanzhe Chen, Yong Wang, Jia Zeng, Mingxuan Yuan, Ting-Chuen Pong, and Huamin Qu. 2019. PlanningVis: A Visual Analytics Approach to Production Planning in Smart Factories. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2019), 579–589.

[52] Thomas Swearingen, Will Drevo, Bennett Cyphers, Alfredo Cuesta-Infante, Arun Ross, and Kalyan Veeramachaneni. 2017. ATM: A distributed, collaborative, scalable system for automated machine learning. In *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 151–162.

[53] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58, 1 (1996), 267–288.

[54] Jarke J Van Wijk and Edward R Van Selow. 1999. Cluster and calendar based visualization of time series data. In *Proceedings 1999 IEEE Symposium on Information Visualization (InfoVis' 99)*. IEEE, 4–9.

[55] Junpeng Wang, Liang Gou, Hao Yang, and Han-Wei Shen. 2018a. Ganviz: A visual analytics approach to understand the adversarial game. *IEEE Transactions on Visualization and Computer Graphics* 24, 6 (2018), 1905–1917.

[56] Qianwen Wang, Yao Ming, Zhihua Jin, Qiaomu Shen, Dongyu Liu, Micah J Smith, Kalyan Veeramachaneni, and Huamin Qu. 2019. ATMSeer: Increasing Transparency and Controllability in Automated Machine Learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 681.

[57] Yong Wang, Hammad Haleem, Conglei Shi, Yanhong Wu, Xun Zhao, Siwei Fu, and Huamin Qu. 2018b. Towards easy comparison of local businesses using online reviews. In *Computer Graphics Forum*, Vol. 37. Wiley Online Library, 63–74.

[58] Marc Weber, Marc Alexa, and Wolfgang Müller. 2001. Visualizing time-series on spirals. In *Infovis*, Vol. 1. 7–14.

[59] M.B. Wilk and R. Gnanadesikan. 1968. Probability plotting methods for the analysis of data. *Biometrika* 55 (1968), 1–17.

[60] Billy M Williams. 2001. Multivariate vehicular traffic flow prediction: evaluation of ARIMAX modeling. *Transportation Research Record* 1776, 1 (2001), 194–200.

[61] Jian Zhao, Fanny Chevalier, and Ravin Balakrishnan. 2011. KronoMiner: using multi-foci navigation for the visual exploration of time-series data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1737–1746.

[62] Xun Zhao, Yanhong Wu, Weiwei Cui, Xinnan Du, Yuan Chen, Yong Wang, Dik Lun Lee, and Huamin Qu. 2017. Skylens: Visual analysis of skyline on multi-dimensional data. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2017), 246–255.