Revista de Contabilidade e Organizações
www.rco.usp.br

DOI: http://dx.doi.org/10.11606/issn.1982-6486.rco.2020.171481

Journal of Accounting and Organizations
www.rco.usp.br

Artificial intelligence applied to small businesses: the use of automatic feature engineering and machine learning for more accurate planning

Inteligência artificial aplicada a pequenas empresas: o uso da engenharia automática de recursos e do aprendizado de máquina para um planejamento mais preciso

Alexandre Moreira Nascimento[a], Vinicius Veloso de Melo[b], Anna Carolina Muller Queiroz[a], Thomas Brashear-Alejandro[cd], Fernando de Souza Meirelles[d]

[a] *Stanford University*
[b] *The Wawanesa Mutual Insurance Company*
[c] *Massachusetts University*
[d] *Fundação Getúlio Vargas*

Keywords
Artificial intelligence.
Automatic feature engineering.
Machine learning.
Small business.
Local business.

Abstract

The purpose of this study is to develop a predictive model that increases the accuracy of business operational planning using data from a small business. By using Machine Learning (ML) techniques feature expansion, resampling, and combination techniques, it was possible to address several existing limitations in the available research. Then, the use of the novel technique of feature engineering allowed us to increase the accuracy of the model by finding 10 new features derived from the original ones and constructed automatically through the nonlinear relationships found between them. Finally, we built a rule-based classifier to predict the store's revenue with high accuracy. The results show the proposed approach open new possibilities for ML research applied to small and medium businesses.

Palavras-chave
Inteligência artificial.
Engenharia automática de recursos.
Aprendizado de máquina.
Pequenas empresas.
Empresas locais.

Resumo

O objetivo deste estudo é desenvolver um modelo preditivo que aumente a precisão do planejamento operacional de negócios usando dados de uma pequena empresa. A partir de técnicas de aprendizado de máquina (AM), são apresentadas estratégias de expansão, reamostragem e combinação que permitiram superar várias das limitações enfrentadas pelas pesquisas conduzidas até então. O estudo adotou uma nova técnica de engenharia de recursos que permitiu aumentar a precisão de um modelo preditivo, encontrando 10 novos recursos derivados dos originais, desenvolvidos automaticamente através das relações não-lineares encontradas entre eles. Por fim, foi criado um classificador com regras para prever, com alta precisão, a receita da pequena empresa. De acordo com os resultados apresentados, a abordagem proposta abre novas possibilidades para a pesquisa sobre a AM aplicada a pequenas e médias empresas.

Article information
Received: June 27, 2020
Approved: September 14, 2020
Published: October 14, 2020

Practical Implications

This study shows how available ML techniques can be used by small businesses even with limited and scarce data to forecast total daily sales revenue. This information can improve their planning precision which can reduce inventory losses and revenue losses due to a lack of inventory to meet demand spikes.

Corresponding author: Tel. +55 (11) 3799-7755
E-mail: alexandremoreiranascimento@alum.mit.edu (A. M. Nascimento); vvdemelo@wawanesa.com (V. V. de Melo); acmq@stanford.edu (A. C. Muller Queiroz); Brashear@isenberg.umass.edu (T. Brashear-Alejandro); fernando.meirelles@fgv.br (F. de S. Meirelles)
Fundação Getulio Vargas, Escola de Administração de Empresas de São Paulo. Av. Nove de Julho, 2029 - Bela Vista, São Paulo/SP - 01313902, Brazil.

1 INTRODUCTION

Uncertainty in sales is a great pain for smaller businesses (Lensink, Van Steen, & Sterken, 2005; Love & Hoey, 1990). A meta-analysis of the contextual factors impacting the business planning-performance relationship in small firms identified a lack of information on sales cycles (Brinckmann, Grichnik, & Kapsa, 2010). Access to sales forecasts seems to be an important factor in small business success (Brinckmann et al., 2010).

Recent advances in machine learning (ML), a class of artificial intelligence (AI) techniques, can be used to help reducing uncertainty in sales. They can do it by forecasting sales based on available data and finding complex relationships between distinct factors and sales. As they are able to model how those factors can influence the sales, those techniques can be used to predict daily sales of smaller businesses, reducing their planning and operational uncertainties. In fact, by understanding the factors that influence sales, the managers can better define policies, strategies, and tactics to influence positively these factors (Moore, 2008).

However, few studies have focused on using advanced ML techniques to predict the daily sales revenue of small businesses, as well as grocery stores and supermarkets. Indeed, a search on Scopus Database using the Boolean search string (("forecast") AND ("supermarket" OR "grocery store") AND ("daily") AND ("demand" OR "revenue" OR "sales")) found only 15 studies. From those, only 8 are related to the topic. However, none of those 8 studies (Aburto & Weber, 2007; Berry, Helman, & West, 2020; Bousqaoui, Achchab, & Tikito, 2019; Deb, 2017; Kolassa, 2013; Slimani, El Farissi, & Achchab, 2017; Slimani, Farissi, & Al-Qualsadi, 2016; J. W. Taylor, 2011) are related to small business and none applied automatic feature engineering..

In fact, there is a gap in the literature on the use of ML to applied in small physical retailers. Since most applications of ML require large datasets with many variables to be able to induce useful models, the practical applicability of those techniques in those businesses becomes very restricted because of their lack of available data. Consequently, small businesses owners still rely on their intuition for most decisions (Culkin & Smith, 2000; Fadahunsi, 2012).

In this context, this study proposes to overcome the data scarcity limitation by using one ML technique for automatic feature engineering recently proposed—Kaizen Programming (KP)—and a creative approach based on the use of publicly secondary data to expand a very scarce exclusive dataset from a small grocery store to predict the future daily sales revenue. By using that information, the store manager can enhance its planning and decision making. This study opens an avenue for research in this area.

This remained of the paper contains four sections and the content is organized as follows. Section 2 provides some background on the transformation in the retail industry which can create opportunities to small players. Section 3 presents details of the methodology adopted into this study. The results are presented in section 4 and the discussion in Section 5. Finally, the conclusion of the study is presented in Section 6.

2 TRANSFORMATION OF RETAIL INDUSTRY – AN OPPORTUNITY FOR SMALL PLAYERS

The phenomenon of digitization and advances in AI are causing the disruption of many traditional industries (Stone et al., 2016; K. Taylor & Hanbury, 2018). One such industry is retail (Gilbert, 2015). Traditional companies with strong brands are struggling to adapt and compete with companies that have been born with digital DNA (Loebbecke & Picot, 2015). Consequently, while traditional chains like Toys 'R Us are closing their doors (Isidore, Wattles, & Kavilanz, 2018) in the United States, Amazon is expanding its presence to physical stores (Soper, 2017).

Large retail companies are falling behind and failing (K. Taylor & Hanbury, 2018), even with their considerable economic and political resources. The calamity also endangers smaller retailers more acutely (Corkery, 2018). In fact, the smaller local retailer has been suffering for years with the difficulties imposed by big manufacturers' and big competitors' pricing and purchasing power respectively (Bowman, 2016). While large networks have bargaining power due to their volumes, they are able to operate on low margins and, therefore, offer low prices. This puts smaller local retailers at a disadvantage, as they require higher margins due to higher overhead, cost structures, and lower sales volume (Draganska, Klapper, & Villas-Boas, 2010).

To compete and survive, some small and medium retailers have been focusing on niches, such as healthy and organic food produced with ingredients from local suppliers (Dagevos, 2016). These niche strategies have a specific focus such as a premium clientele (Gil, Gracia, & Sanchez, 2000; Thompson, 1998; Van Doorn & Verhoef, 2011), focus on quality over the price (Doward, 2017), and leverage exclusivity, which can boost margins, especially with small-scale sales and turnover. On the other hand, often, as niches prove to be viable through revenue growth and regional expansion, attentive big retailers move to grab another slice of the market (Tu, 2016). In the US, Whole Foods is an example of a retail store that grew through acquisition within a niche, and it was acquired by Amazon (Wingfield & de la Merced, 2017), which decided to enter the physical retail business and offer healthy, high-margin products (Leswing, 2017). Therefore, small and medium businesses struggle to sustain competitive advantage even in niche markets, especially in the retail and grocery store markets.

Historically traditional large retailers have been having much more easy access to information technology innovations when compared to small and medium retailers because of the size of the investments needed to acquire and maintain the required computational infrastructure to adopt them. Therefore, those large retailers could sustain some competitive advantage even being slower to adapt to technological changes (Loebbecke & Picot, 2015). Small and medium retailers are traditionally faster and subjected to more constant adaptation (Li, Su, Liu, & Li, 2011), which are crucial factors for their growth and survival. However, their access to innovative technology have been historically limited by their shortage of resources to invest.

However, the digital revolution, supported by cheap and available connectivity, cloud computing, open source tools and advancements in AI and ML, have democratized access to powerful innovative technologies (Dewhurst & Willmott, 2014; Müller & Bostrom, 2016). The impact of this phenomenon, seen in an increasing number of financial technology (fintech) startups, is to reduce barriers to accessing technological tools (e.g., planning, decision making, automation, and optimization previously only accessible to large firms) (Dapp & Slomka, 2015). Thus, smaller businesses, even those without a history of innovation, can have now an opportunity to leverage these technologies to reduce their disadvantages compared to large companies and potentially create competitive advantages through their speed of adaptation and ease of incorporating change (Banks, 2013; Christensen & Bower, 1996; Cohen & Klepper, 1996; Davenport & Bibby, 1999; Jones, 2004). Therefore, in this context of transformation, small and medium retailers can now access cost effective and widely available information technologies. Among them, techniques based on AI can bring competitive advantage to those businesses (Burns, 2016; Gordon & Key, 1987; McFarlane, 1984) since they can be used to predict sales and improve the accuracy of planning and operational optimization on several fronts.

Indeed, uncertainty in sales is one of the greatest pains of small business (Lensink *et al.*, 2005) and a risk factor in the food business (Deb, 2017; Love & Hoey, 1990). Uncertainty in food business sales can result in daily losses due to over or under stocking (Deb, 2017; Ha, 1997). Over stocking in stores with own production of natural and healthy products can yield to higher losses. This is because the lack of preservatives and chemical additives in those products shortens their shelf life (Bonti-Ankomah & Yiridoe, 2006). On the other hand, under stocking can lead to sales losses due to stockouts (Deb, 2017; Ha, 1997), thereby reducing the revenue. Thus, these stores seek low inventories, frequent deliveries by suppliers, and adjustments based on the intuition and perception of the attendants, hence avoiding the losses of their products (Caspi, Pelletier, Harnack, Erickson, & Laska, 2016; Dunkley, Helling, & Sawicki, 2004).

The resulting subjective annoyance is smaller than the potential for unpredictable lost sales. This policy reduces risk, but it increases the operational overhead of handling constant orders (since they cannot be preprogrammed due to constant fine tuning) (Andreyeva, Middleton, Long, Luedicke, & Schwartz, 2011). The cost of constant deliveries does not eliminate losses or eliminate the loss of revenue due to lost sales. In this scenario, the owners of these establishments operate under sub-optimal conditions and become averse to offering new products (Andreyeva *et al.*, 2011).

Understanding the factors that influence sales is critical for retail owners. It allows them to define policies, strategies, and tactics and to work on variables that impact these factors (Moore, 2008). For example, estimating the average number of in-store visits allows for the optimization of operational capacity planning, such for the required number of attendants and cashiers, inventory optimization, cash and debt planning, security, and, in the case of companies that do deliver, for the associated logistical dimensioning issues (Lee & Whang, 2000; Moore, 2008). However, small businesses owners usually use intuition for most decisions (Culkin & Smith, 2000; Fadahunsi, 2012).

3 METHODOLOGY

This section describes the methodology used in this research. Figure 1 shows a diagram of protocol used in this work. Initially, a research question was defined. Then, the data were prepared. Right after, it was followed by an analysis using an innovative Feature Engineering technique. A resampling technique was employed due to the imbalance of the data sample. Finally, the typical process of inducing a ML classifier was used, which encompasses the ML training and testing. Finally, the results were reported.

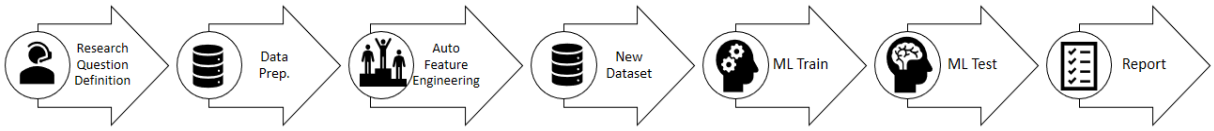

Figure 1. Research process
Source: authors.

3.1 Research question definition

We defined the research question to guide the application of the techniques in order to provide predictive data that could help managers of small companies to enhance the accuracy of their planning. The defined research question was: *"How is it possible predict the daily sales revenue of a small grocery store using the scarce information available?"*

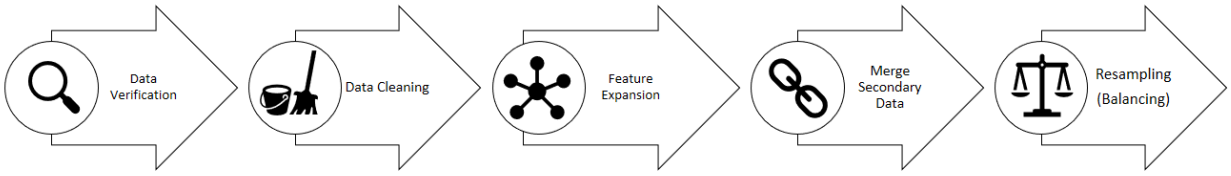

Figure 2. Data preparation
Source: authors.

3.2 Data preparation

The data obtained were the daily sales revenue of one store over 1 year beginning in August 1, 2016 and ending July 31, 2017. Several procedures were used to prepare the data as shown in Figure 2. The first, data verification, addressed the formatting of the data and the existence of missing values. Given the data consisted of only the date and the corresponding daily sales revenue, it was necessary to perform a treatment to derive new features manually from the existing attributes. This treatment maintains important characteristics and provides tips for using ML models.

A new attribute—the day of the week—was generated and added in the dataset. This attribute was added in numeric format. Then, the date was partitioned by day of the month and month so that the original attribute was replaced by two new numeric attributes. In this way, we could maintain useful weekly and monthly seasonal information. Finally, to ensure that information about the sequence of the days was not lost, a numerical sequential field was added to store the sequence of each of the days throughout the year.

Initially it was intended to create a model that could predict the daily revenue for a future day with high precision. That is, the model should be able to tell the store manager, for example, next Monday, the daily sales revenue will be R$ 7325,00, and this value when compared to the real revenue would present a low error rate. However, after many attempts using ML regression-based techniques, it was noticed that it was not possible to reach an adequate precision, due to the low amount of data. Therefore, aiming to overcome this limitation, it was decided to use ML classification models to forecast the level of daily sales revenue. As validated with the store manager, this would still be helpful since it would narrow down the range of the expected daily sales, which would provide a more precise target for planning when compared to the management intuition.

For this reason, a new attribute—daily sales revenue level—has been defined in order to replace daily sales revenue so the model it would work with discrete daily sales revenue intervals as its target variable. This attribute was obtained by partitioning the historical daily sales revenue range [1.12e+03, 9.28e+03] into 5 levels defined by 5 distinct ranges of same size (1.63e+03): very low {a = (1.12e+03, 2.75e+03]}; low {b = (2.75e+03, 4.38e+03]}; medium {c = (4.38e+03, 6.01e+03]}; high {d = (6.01e+03, 7.64e+03]}; and very high {e = (7.64e+03, 9.28e+03]}. It is noteworthy that a reduced number of levels, such as 3, for example, would increase the model's accuracy, but it would not provide and adequate resolution for the store manager usage. On the other hand, a higher number of levels would reduce the model's accuracy, providing false expectations or certainties to the store manager's plan, what could potentially result into undesired losses.

Subsequently, given the small number of attributes to be used for a model that could explain the dependent variable (prediction of the daily sales revenue level), the data were supplemented with secondary climatic data. Through a historical database of climate information (https://www.wunderground.com), we obtained the following attributes: the daily information of temperature (maximum and minimum), relative humidity of the air, and occurrence of a climatic event (presence of wind, thunderstorm, rain etc.). With this, the climatic data obtained were combined with the primary data, obtaining a new database. Finally, another feature was added: a flag indicating whether the day was a holiday or not. Table 1 shows the features (variables) included in the final version of the dataset used as input for the automatic feature engineering step. The model defined as the target of the analysis is given by:

DAILY-REVENUE-LEVEL = f (SEQUENCE, DAYOFWEEK, DAYOFMONTH, MONTH, HOLIDAY, TEMPMAX, TEMPMIN, HUMIDITY, CLIMATE-EVENT)

Table 1. Dataset features (variables) before automatic feature engineering

Variable	Description
SEQUENCE	a unique and sequential number from 1 to 312 to keep the information of the sequence of each day
DAYOFWEEK	number corresponding to the day of the week (1 for Sunday through 7 for Saturday)
DAYOFMONTH	the day of the month (1-31)
MONTH	the month of the year (1-12)
HOLIDAY	0 indicates that the day is not a holiday and 1 indicates that it is a holiday
TEMPMAX	maximum temperature of the day in Celsius
TEMPMIN	minimum temperature of the day in Celsius
HUMIDITY	percentage of relative humidity
CLIMATE-EVENT	0 indicates no occurrence of climate event in the day, such as rain, thunderstorm etc.; 1 indicates one or more climate event occurrences in the day

Source: authors.

An analysis of the constructed dataset showed an even more pronounced problem because of the small amount of data available: an unbalanced dataset with respect to under-represented minority classes. The number of observations for each class is a=36, b=109, c=137, d=28, and e=2. Thus, a resampling technique was used to increase the examples of the minority classes in order to balance the dataset (Albisua *et al.*, 2013; Batuwita & Palade, 2010; Estabrooks, Jo, & Japkowicz, 2004; Ramentol *et al.*, 2012). This strategy was used only to force the ML algorithm to include minority classes in the model.

3.3 Automatic feature engineering

Classification techniques, such as decision tree algorithms, for example, seek rules based on a range of the attributes so that those rules can split the multidimensional space into different regions or classes. Considering the decision tree algorithms, for example, this is achieved by splitting each decision node (the internal nodes of the decision tree) attribute according to some criterion, such as the split that gives the minimum entropy in terms of class distribution (the lower the entropy, the more separated the classes are). A sequence of splits can be converted into a decision rule. Figure 3 illustrates a hypothetical decision tree with the set of rules, with the best fit in splitting the plane XY based on x and y, to separate the two different classes (x and red filled circle).

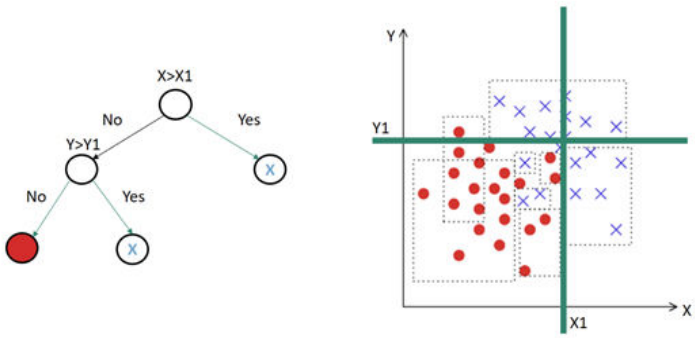

Figure 3. Data classification on a decision tree
Source: authors.

A clear characteristic of this approach is that the attributes are evaluated independently, resulting in splits along the axes, ignoring possible relationships among variables. To deal with such an issue, one can perform manual or automatic feature engineering to combine the attributes. The Automatic Feature Engineering technique used in this research is called Kaizen Programming (KP) (De Melo, 2014).

KP is an iterative technique based on a computational abstraction of the Kaizen methodology with the Plan-Do-Check-Act (PDCA) cycle. KP divides a problem into parts for investigation. Each part can be evaluated separately (a partial solution), improved, and put together in a complete solution; then, each part's contribution (quality) can be assessed along with the quality of the complete solution. Thus, one can verify which partial solutions were important and establish a ranking. Excerpts of the solution that do not generate relevant contributions can be discarded, while the others are maintained in the improvement process.

KP is a hybrid approach combining a global search algorithm, such as an Evolutionary Algorithm, and an efficient local search technique to build the actual model. In KP, global and local search techniques work differently from traditional approaches. It means that the main algorithm in KP is the local search (which builds the complete solution), while the global search only tries to identify promising regions (the partial solutions) of the search space.

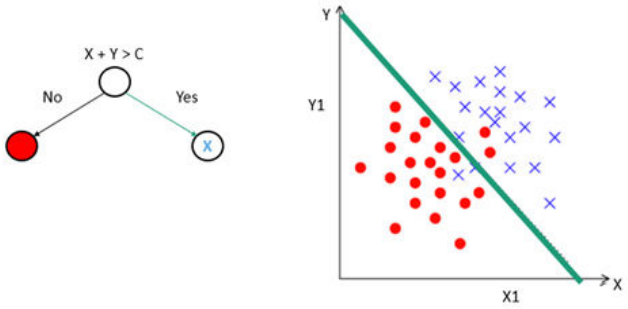

Figure 4. Data classification rules based feature combination
Source: authors.

In summary, the global search algorithm looks for useful relationships among the attributes of the dataset, while the local search algorithm builds the complete model using the new relationships. The global search used here is an evolutionary computation technique called Genetic Programming (GP) that evolves solutions (individuals) whose genes are mathematical expressions. For instance, suppose the dataset has attributes X and Y, and the available expressions are +, -, *, and /, a possible individual (ind) generated by GP can be ind = X + Y. GP has a population of individuals that evolve through many generations, optimizing a certain objective function, i.e., maximizing the importance of each new attribute in a classifier. The classification technique, in our case, is a decision tree, which provides the importance of each attribute used to build the model. The bigger the importance, the better the attribute; consequently, the better the individual in the population. This way, GP can evolve better features that will result in better decision trees.

In other words, KP finds formulas combining those attributes that can wisely split the space and bring about better classification results. By finding those formulas, there is an information gain (Entropy Gain = Entropy Before – Entropy After), which can provide additional features (feature expansion) or can substitute many features with one that is based on a formula that combines them (feature reduction). Figure 4 shows an example where the decision tree uses the new attribute discovered by GP and finds the best split value (C) to separate the classes. One can observe the solution is not only better than in Figure 3, but smaller.

3.4 Building new features

There are two steps involved in the approach evaluated in this study. First, KP has to discover features that must be significant to the classifier. Then, we create a new dataset by appending the new features after the original ones. An important aspect of this research is that the final objective is to predict the level of the daily sales revenue through a classification task. However, the original dependent variable is a real number. Thus, we investigated a strategy wherein KP evolves new features in a regression task using an ordinary least squares algorithm. Later, a classification algorithm uses these new features with a discretized dependent variable, resulting in the desired classification task.

For KP, we performed 50 independent runs to evolve 10 new features. The available functions to create a new feature were +, -, *, /, log, sqrt, abs (modulo), and 1/x, where x can be a value or an expression. In the Check step, KP generates five new features from each of the 10 current features (the standard). Then, KP removes the features with a co-correlation of > 0.9 (keeping the one that appeared first in the dataset) and uses ordinary least squares algorithm to create a model and calculate the importance of each uncorrelated feature. Finally, the 10 best features are chosen for the new trial model. If the quality, measured by the mean squared error, of the model using the new features is better than that of the standard model, then the latter is updated. This cycle is repeated 2,000 times for each run. At the end of this process, we selected the set of features of the run that resulted in the best final quality.

3.5 Classification technique

Classification employed the PART algorithm in the Weka ML tool version 3.6.15. We used, PART, a rule learner that creates a list of decision rules by employing a separate-and-conquer approach, building a partial decision tree at each iteration and turning the "best" leaf into a rule. We chose a rule learner because rules usually result in white-boxes while more sophisticated algorithms, such as artificial neural networks, result in black-boxes. Although artificial neural networks are usually able to achieve a better prediction quality, as a black-box approach they are not able to offer human comprehensible insights on the relationship between the dependent variable and the independent ones. On the other hand, PART builds rules based on the features discovered by KP. Thus, while the rules are a white-box, some of the new features may not be interpretable, making the final result a grey-box. Nevertheless, there is a useful relationship revealed among the original features when they belong to the same new feature.

3.6 Training the classifier

Given that the classification dataset is unbalanced, we employed the supervised resampling filter in Weka to balance the dataset in order to force the generation of rules to cover all classes. Otherwise, the small classes would be “absorbed” by the larger classes, meaning that the technique prefers a simpler model for better generalization instead of creating rules to correctly classify just a few observations (specialization). The configuration, chosen after an empirical analysis of the distribution, was 0.95 of bias and 250% of sample size. No other filter was used. Regarding the classification technique, we used PART with its default configuration in the tool.

3.7 Testing the classifier

In our approach, the augmented dataset was used in the training, while the original (unbalanced) dataset was used to test the resulting classifier. This is not the same as using the training set for testing, which could provide a perfect prediction (Farhadi, 2018).

4 RESULTS

The best features discovered by KP, ordered from the most to least important features, are shown below. Notably, there are structures repeated in different features, such as 1/log(log(<F>)) and log(<F>/log(<F>)), where <F> refers to some of the original features. Although similar, the features have quite different distributions, as can be seen in Figure 5.

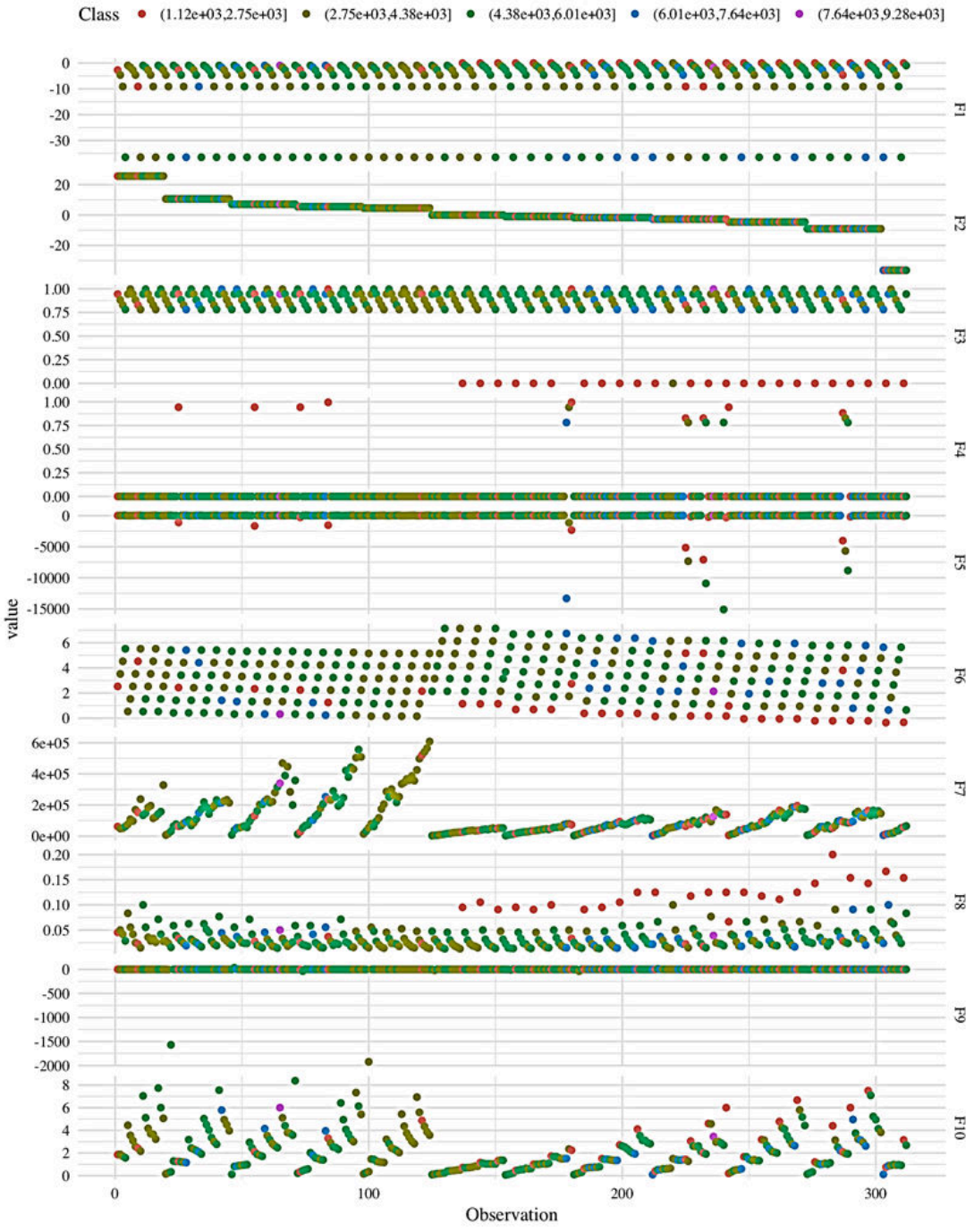

Figure 5. Plot of each new feature for the 312 observations in the dataset
Source: authors.
Note: verify Apendix for more information

The result of the test set using PART is shown in Table 2 and the confusion matrix is shown in Table 3. It can be seen that contrary to what was observed in the preliminary tests, the results point to a good class differentiation and a marked degree of success.

Table 2. Results of PART using default configuration

Correctly Classified Instances	253 (81.09%)
Incorrectly Classified Instances	59 (18.91%)
Kappa statistic	0.7249
Mean absolute error	0.0803
Root mean squared error	0.2687
Relative absolute error	25.09%
Root relative squared error	67.15%
Total Number of Instances	312

Source: authors.

Table 3. Confusion matrix

a	b	c	d	e	← classified as
36	0	0	0	0	a = (1.12e+03,2.75e+03]
3	89	12	4	1	b = (2.75e+03,4.38e+03]
1	28	98	9	1	c = (4.38e+03,6.01e+03]
0	0	0	28	0	d = (6.01e+03,7.64e+03]
0	0	0	0	2	e = (7.64e+03,9.28e+03]

Source: authors.

Note: Black – right classification; Gray – wrong classification

Table 4 shows the precision of the classification performed for each class (daily sales revenue level). Classes a, d, e (very low, high, and very high) had 100% of the classifications done correctly. Classes b and c (low and medium) had 82% and 72%, respectively, of the classifications done correctly. Of 312 days, 59 (18.9%) were incorrectly classified.

Table 4. Precision of the classification per classes

Classes	Precision
a = (1.12e+03,2.75e+03]	100%
b = (2.75e+03,4.38e+03]	82%
c = (4.38e+03,6.01e+03]	72%
d = (6.01e+03,7.64e+03]	100%
e = (7.64e+03,9.28e+03]	100%

Source: authors.

5 DISCUSSION

Applying the method used in this study demonstrated the level of daily sales revenue could be correctly forecasted in 81% of the cases during the execution of the test dataset. Therefore, the model that was based on the features (formulas) discovered by KP correctly forecasted revenue levels in 253 of 312 days. By knowing the daily sales revenue level in advance, the manager can estimate a forecast for the number of customers on a given day by simply dividing the low-end and upper-end of the revenue range in the level (or the average of both) as forecasted by the average ticket (around R$ 50,00 per transaction).

The number of customers visiting the store per day can be used as an input for planning the production of own-production products and the needed inventory for third-party products. The manager can calculate the average number of each product (own-production or third-party) sold per customer per day in a period of previous days and use those ratios to calculate the expected sales of those products on the following days.

The proposed model provided more explanation for the extremes than the intermediate daily sales revenue levels. However, it was able to reach a reasonable precision level (>70%) for the intermediate levels as well. As the extremes are the lowest and highest levels of daily sales revenue, the model can help to reduce large revenue losses and carrying costs by better predicting daily and weekly inventory needs, refined schedules for production, and employees. The model provides a powerful tool for increasing small business managers' abilities to recognize and mitigate extreme situations.

An important aspect of our study relates to the complexity of the features discovered by KP. There is trade-off in the technique between the understandability of the discovered formulas and the precision on the forecast. Therefore, aiming to maximize the precision of the classifications, KP was configured to the high level of complexity of the generated formulas. The advantage is that a good result was achieved in the classification, while the disadvantage is that the formulas are hard to understand and to be used rationally by managers of small businesses. On the other hand, one could decide to use only arithmetical operations and limit the expression lengths to obtain simpler features.

However, even with their high complexity, those formulas can serve as useful tools for the manager. In fact, they can be programmed into a spreadsheet or a simple computer program. Then, the manager can simply provide the input values (date, weather forecast info, and if it is holiday) and have the formulas calculated automatically. Then, the classification algorithm can forecast the daily sales revenue level.

Consequently, managers can use the tool as a simulator to understand the impact of different factors on daily sales revenue levels. The forecasted information can be used to support management's fine-tuning of various operational aspects, such as the production of its own products, order sizes and frequency, employee schedules, and allocations. For example, employees can be allocated for organizing duties on days with low shoppers' visits and for customer service on high-demand days. Additionally, managers could influence the number of store visits on days when daily sales are expected to be low. For example, by anticipating that low temperatures will reduce daily sales, the store can send e-mails to customers days before offering free hot chocolate on slow days, or emails could be sent for very warm days with high relative humidity levels offering free organic ice cream or juice.

6 CONCLUSION

This study innovates on the use of a new daily revenue prediction method that uses an exclusive dataset and is applied to a real small business. It is intended to show how available sophisticated ML techniques can be used by small businesses to enhance their performance. The tools allow management to increase the precision of planning, minimizing inventory losses, on the one hand, and reducing revenue losses due to a lack of inventory to meet demand spikes. The results of the analysis show how a small firm, even with limited and scarce data, can use these techniques to forecast total daily sales revenue, which can be used to derive the number of shoppers from it. The impact of data scarcity and limitation, as showed in the current study, can be reduced by combining external datasets (secondary data, such as weather) and then performing manual feature expansions, automatic feature engineering, and data resampling to balance the dataset. As shown, the forecasted information can be used to support management in fine tuning many operational aspects, such as the production of the businesses' own products, anticipating orders size and supplier frequency, and delivery and employee schedules. Although the dataset used was from a small local grocery store focused on organics and healthy food, the proposed method can be used for restaurants and other small businesses.

The results provide evidence that the innovative combination of methods and techniques employed is promising. In fact, the test performed correctly forecasted daily sales revenue levels on 253 days (81%) of the test period (312 days). They also demonstrate that sophisticated ML techniques are within the reach of small businesses even with data scarcity and limitations and can help them to capture the value offered by AI techniques. KP played an important role by uncovering new features that could enhance the precision of classification. This demonstrated that KP could be a powerful tool for business data analysis.

The present study has limitations that the multidisciplinary research team worked to address through the selection of appropriate data science techniques. Among these limitations are its small dataset, with only 312 samples, the sub-representability of some daily revenue bands, causing an unbalance in the database, and, finally, the few attributes included in the original database. Hence, it is recommended that studies without these limitations be performed to amplify the results reported here.

Finally, the good use of AI techniques by small firms offers a fertile field for research with potential social impact by enabling small firms to become more competitive. Further research is recommended for addressing the limitations reported here as well as for broadening research using other secondary data sources or increasing the collection of primary data through simple techniques that encourage customers to provide more information to the company. Finally, it is recommended that similar research be conducted with small businesses in different sectors to increase the repertoire of available data on the subject.

REFERENCES

Aburto, L., & Weber, R. (2007). A sequential hybrid forecasting system for demand prediction. *Lecture Notes in Computer Science*, 4571, 518–532. DOI: https://doi.org/10.1007/978-3-540-73499-4_39

Albisua, I., Arbelaitz, O., Gurrutxaga, I., Lasarguren, A., Muguerza, J., & Pérez, J. M. (2013). The quest for the optimal class distribution: an approach for enhancing the effectiveness of learning via resampling methods for imbalanced data sets. *Progress in Artificial Intelligence*, 2(1), 45–63. DOI: https://doi.org/10.1007/s13748-012-0034-6

Andreyeva, T., Middleton, A. E., Long, M. W., Luedicke, J., & Schwartz, M. B. (2011). Food retailer practices, attitudes and beliefs about the supply of healthy foods. *Public Health Nutrition*, 14(6), 1024–1031. DOI: https://doi.org/10.1017/S1368980011000061

Banks, G. P. (2013). *Exploring small-business change and strategic adaptation in an evolving economic paradigm*. Doctoral dissertatio. Walden University.

Batuwita, R., & Palade, V. (2010). Efficient resampling methods for training support vector machines with imbalanced datasets. In: *The 2010 International Joint Conference on Neural Networks* (IJCNN) (pp. 1–8).

Berry, L. R., Helman, P., & West, M. (2020). Probabilistic forecasting of heterogeneous consumer transaction–sales time series. *International Journal of Forecasting*, 36(2), 552–569. DOI: https://doi.org/10.1016/j.ijforecast.2019.07.007

Bonti-Ankomah, S., & Yiridoe, E. K. (2006). Organic and conventional food: a literature review of the economics of consumer perceptions and preferences. *Organic Agriculture Centre of Canada*, 59, 1–40.

Bousqaoui, H., Achchab, S., & Tikito, K. (2019). Machine learning applications in supply chains: Long short-term memory for demand forecasting. *Lecture Notes in Networks and Systems*, 49, 301–317. DOI: https://doi.org/10.1007/978-3-319-97719-5_19

Bowman, J. (2016). *Walmart's neighborhood market is crushing the competition*. Business Insider. Retrieved May 15, 2018, from: http://www.businessinsider.com/walmarts-neighborhood-market-is-crushing-the-competition-2016-8

Brinckmann, J., Grichnik, D., & Kapsa, D. (2010). Should entrepreneurs plan or just storm the castle? A meta-analysis on contextual factors impacting the business planning--performance relationship in small firms. *Journal of Business Venturing*, 25(1), 24–40. DOI: https://doi.org/10.1016/j.jbusvent.2008.10.007

Burns, P. (2016). *Entrepreneurship and small business*. Palgrave Macmillan Limited.

Caspi, C. E., Pelletier, J. E., Harnack, L., Erickson, D. J., & Laska, M. N. (2016). Differences in healthy food supply and stocking practices between small grocery stores, gas-marts, pharmacies and dollar stores. *Public Health Nutrition*, 19(3), 540–547. DOI: https://doi.org/10.1017/S1368980015002724

Christensen, C. M., & Bower, J. L. (1996). Customer power, strategic investment, and the failure of leading firms. *Strategic Management Journal*, 197–218.

Cohen, W. M., & Klepper, S. (1996). Firm size and the nature of innovation within industries: the case of process and product R&D. *The Review of Economics and Statistics*, 232–243.

Corkery, M. (2018). *Grocery Wars Turn Small Chains Into Battlefield Casualties*. The New York Times. Retrieved May 15, 2018, from: https://www.nytimes.com/2018/03/26/business/grocery-wars-small-chains.html

Culkin, N., & Smith, D. (2000). An emotional business: a guide to understanding the motivations of small business decision takers. *Qualitative Market Research: An International Journal*, 3(3), 145–157. DOI: https://doi.org/10.1108/13522750010333898

Dagevos, H. (2016). *Beyond the Marketing Mix: Modern Food Marketing and the Future of Organic Food Consumption*. In: The Crisis of Food Brands: Sustaining Safe, Innovative and Competitive Food Supply, 255.

Dapp, T., & Slomka, L. (2015). Fintech reloaded - Traditional banks as digital ecosystems. *Publication of the German Original*, from: https://www.deutschebank.nl/nl/docs/Fintech_reloaded_Traditional_banks_as_digital_ecosystems.pdf

Davenport, S., & Bibby, D. (1999). Rethinking a national innovation system: The small country as' SME'. *Technology Analysis & Strategic Management*, 11(3), 431–462.

De Melo, V. V. (2014). Kaizen Programming. In *Proceedings of the 2014 Conference on Genetic and Evolutionary Computation* (pp. 895–902). New York, NY, USA: ACM. DOI: https://doi.org/10.1145/2576768.2598264

Deb, S. (2017). Analytical ideas to improve daily demand forecasts: A case study. *Lecture Notes in Computer Science*, 10192, 23–32. DOI: https://doi.org/10.1007/978-3-319-54430-4_3

Dewhurst, M., & Willmott, P. (2014). Manager and machine: The new leadership equation. *McKinsey Quarterly*, 4, 1–8.

Doward, J. (2017). *Organic food sales soar as shoppers put quality before price* | Environment | The Guardian. Retrieved May 15, 2018, from: https://www.theguardian.com/environment/2017/feb/19/sales-of-organic-food-soar-fruit-vegetables-supermarkets

Draganska, M., Klapper, D., & Villas-Boas, S. B. (2010). A larger slice or a larger pie? An empirical investigation of bargaining power in the distribution channel. *Marketing Science*, 29(1), 57–74. DOI: https://doi.org/10.1287/mksc.1080.0472

Dunkley, B., Helling, A., & Sawicki, D. S. (2004). Accessibility versus scale: examining the tradeoffs in grocery stores. *Journal of Planning Education and Research*, 23(4), 387–401. DOI: https://doi.org/10.1177/0739456X04264890

Estabrooks, A., Jo, T., & Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence*, 20(1), 18–36. DOI: https://doi.org/10.1111/j.0824-7935.2004.t01-1-00228.x

Fadahunsi, A. (2012). The growth of small businesses: Towards a research agenda. *American Journal of Economics and Business Administration*, 4(1), 105. DOI: https://doi.org/10.3844/ajebasp.2012.105.115

Farhadi, H. (2018). *Machine Learning: Advanced Techniques and Emerging Applications*. BoD--Books on Demand.

Gil, J. M., Gracia, A., & Sanchez, M. (2000). Market segmentation and willingness to pay for organic products in Spain. *The International Food and Agribusiness Management Review*, 3(2), 207–226. DOI: https://doi.org/10.1016/S1096-7508(01)00040-4

Gilbert, R. J. (2015). E-books: A tale of digital disruption. *Journal of Economic Perspectives*, 29(3), 165–184. DOI: https://doi.org/10.1257/jep.29.3.165

Gordon, W. L., & Key, J. R. (1987). Artificial intelligence in support of small business information needs. *Journal of Systems Management*, 38(1), 24.

Ha, A. Y. (1997). Inventory rationing in a make-to-stock production system with several demand classes and lost sales. *Management Science*, 43(8), 1093–1103.

Isidore, C., Wattles, J., & Kavilanz, P. (2018). *Toys "R" Us will close or sell all US stores*. Retrieved May 15, 2018, from: http://money.cnn.com/2018/03/14/news/companies/toys-r-us-closing-stores/index.html

Jones, C. (2004). An alternative view of small firm adaptation. *Journal of Small Business and Enterprise Development*, 11(3), 362-370. DOI: https://doi.org/10.1108/14626000410551618

Kolassa, S. (2013). Forecasting and optimisation for big data: Lessons from the retail business. In OR55 Keynotes and Extended Abstracts - *55th Conference of the Operational Research Society*, 33–35. Retrieved from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84887836974&partnerID=40&md5=4a2ff7c1644d89c6f3d3df3073c95d5f

Lee, H. L., & Whang, S. (2000). Information sharing in a supply chain. *International Journal of Manufacturing Technology and Management*, 1(1), 79–93. DOI: https://doi.org/10.1016/j.proeng.2012.06.258

Lensink, R., Van Steen, P., & Sterken, E. (2005). Uncertainty and Growth of the Firm. *Small Business Economics,* 24(4), 381–391. DOI: https://doi.org/10.1007/s11187-005-7121-z

Leswing, K. (2017). *Amazon Is Buying Whole Foods-Here's Amazon's Vision for the Grocery Store of the Future*. Business Insider, from: https://www.businessinsider.com/amazon-go-grocery-store-future-photos-video-2017-6

Li, Y., Su, Z., Liu, Y., & Li, M. (2011). Fast adaptation, strategic flexibility and entrepreneurial roles. Chinese *Management Studies*, 5(3), 256–271. DOI: https://doi.org/10.1108/17506141111163354

Loebbecke, C., & Picot, A. (2015). Reflections on societal and business model transformation arising from digitization and big data analytics: A research agenda. *Journal of Strategic Information Systems*, 24(3), 149–157. DOI: https://doi.org/10.1016/j.jsis.2015.08.002

Love, R. R., & Hoey, J. M. (1990). Management science improves fast-food operations. *Interfaces*, 20(2), 21–29.

McFarlane, F. W. (1984). *Information technology changes the way you compete*. Harvard Business Review, Reprint Service.

Moore, C. W. (2008). *Managing small business: An entrepreneurial emphasis*. Cengage Learning EMEA.

Müller, V. C., & Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. In *Fundamental issues of artificial intelligence* (pp. 555–572). Springer.

Ramentol, E., Verbiest, N., Bello, R., Caballero, Y., Cornelis, C., & Herrera, F. (2012). SMOTE-FRST: a new resampling method using fuzzy rough set theory. In *Uncertainty Modeling in Knowledge Engineering and Decision Making* (pp. 800–805). World Scientific.

Slimani, I., El Farissi, I., & Achchab, S. (2017). Configuration and implementation of a daily artificial neural network-based forecasting system using real supermarket data. *International Journal of Logistics Systems and Management*, 28(2), 144–163. DOI: https://doi.org/10.1504/IJLSM.2017.086345

Slimani, I., Farissi, I. E., & Al-Qualsadi, S. A. (2016). Configuration of daily demand predicting system based on neural networks. In Proceedings of the *3rd IEEE International Conference on Logistics Operations Management*. DOI: https://doi.org/10.1109/GOL.2016.7731709

Soper, T. (2017). *Amazon reports $1.3B in physical store sales, breaking out brick-and-mortar business for first time, still dwarfed by $26.4B online sales*. GeekWire. Retrieved May 15, 2018, from: https://www.geekwire.com/2017/amazon-adds-physical-stores-segment-earnings-report-expands-brick-mortar-footprint/

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., … Horvitz, E. (2016). *Artificial intelligence and life in 2030: One hundred year study on artificial intelligence*. Standford University, from: https://ai100.stanford.edu/sites/g/files/sbiybj9861/f/ai_100_report_0831fnl.pdf

Taylor, J. W. (2011). Multi-item sales forecasting with total and split exponential smoothing. *Journal of the Operational Research Society*, 62(3), 555–563. DOI: https://doi.org/10.1057/jors.2010.95

Taylor, K., & Hanbury, M. (2018). *Amazon is threatening these 8 industries*. Business Insider. Retrieved May 15, 2018, from: http://www.businessinsider.com/amazon-is-killing-these-7-companies-2017-7#department-stores-3

Thompson, G. D. (1998). Consumer demand for organic foods: what we know and what we need to know. *American Journal of Agricultural Economics*, 80(5), 1113–1118. DOI: https://doi.org/10.2307/1244214

Tu, J. I. (2016). *Costco gets creative to meet shoppers' huge appetite for organics*. The Seattle Times. Retrieved May 15, 2018, from: https://www.seattletimes.com/business/retail/costco-gets-creative-to-meet-shoppers-huge-appetite-for-organics/

Van Doorn, J., & Verhoef, P. C. (2011). Willingness to pay for organic products: Differences between virtue and vice foods. *International Journal of Research in Marketing*, 28(3), 167–180. DOI: https://doi.org/10.1016/j.ijresmar.2011.02.005

Wingfield, N., & de la Merced, M. (2017). *Amazon to buy Whole Foods for 13.4 billion*. The New York Times.

How to cite this paper

Nascimento, A. M.; de Melo, V. V.; Muller Queiroz, A. C.; Brashear-Alejandro, T.; & Meirelles, F. de S. (2020). Artificial intelligence applied to small businesses: the use of automatic feature engineering and machine learning for more accurate planning. *Revista de Contabilidade e Organizações*, 14:e171481. DOI: http://dx.doi.org/10.11606/issn.1982-6486.rco.2020.171481

Appendix

$$F1 = \frac{1}{log\left(log\left(\sqrt{DAYOFWEEK}\right)\right)}$$

$$F2 = \frac{1}{log\left(log\left(\sqrt{MONTH}\right)\right)}$$

$$F3 = \frac{1}{log\left(\frac{DAYOFWEEK}{log(DAYOFWEEK)}\right)}$$

$$F4 = \frac{HOLIDAY}{log\left(\frac{DAYOFWEEK}{log(DAYOFWEEK)}\right)}$$

$$F5 = \frac{1}{2}\left[DAYOFWEEK + \left(DAYOFWEEK - HOLIDAY * \left(\frac{DAYOFMONTH + \frac{C(LIMATEEVENT + AbsDAYOFMONTH * DAYOFWEEK * HUMIDITY)}{2}}{2}\right)\right)\right]$$

$$F6 = DAYOFWEEK - log\left(\frac{1}{2} * \left(MONTH + \left(\sqrt{\frac{1}{HOLIDAY + \sqrt{log(abs(TEMPMAX))}}}\right)\right)\right)$$

$$F7 = abs\left(\frac{\left(\frac{log(DAYOFWEEK)}{DAYOFMONTH} - absTEMPMIN - HOLIDAY\right) * HUMIDITY * MONTH}{DAYOFMONTH}\right)$$

$$F8 = \frac{1}{DAYOFWEEK * \frac{TEMPMIN + CLIMATEEVENT}{2}}$$

$$F9 = \frac{1}{log\left(\frac{DAYOFMONTH}{\sqrt{\frac{\left(TEMPMIN + \left(\frac{TEMPMIN}{TEMPMAX^2}\right)\right)}{2}}}\right)}$$

$$F10 = \frac{DAYOFMONTH}{\left(DAYOFWEEK + abs\left(\frac{1}{abs\left(\frac{MONTH}{TEMPMIN}\right)}\right)\right)}$$