Association for Information Systems
AIS Electronic Library (AISeL)

Research Papers | ECIS 2019 Proceedings

5-15-2019

EXPLAINING CUSTOMER ACTIVATION WITH DEEP ATTENTION MODELS

Koen Weterings
Open University, koen.weterings@apg.nl

Stefano Bromuri
Open University, stefano.bromuri@ou.nl

Marko van Eekelen
Open University, marko.vaneekelen@ou.nl

EXPLAINING CUSTOMER ACTIVATION WITH DEEP ATTENTION MODELS

Complete Research

Weterings, Koen, Open University of the Netherlands, Heerlen, the Netherlands, koen.weterings@ou.nl

Bromuri, Stefano, Open University of the Netherlands, Heerlen, the Netherlands, stefano.bromuri@ou.nl

Eekelen, Marko van, Open University of the Netherlands, Heerlen, the Netherlands and Radboud University, Nijmegen, the Netherlands, marko.vaneekelen@ou.nl

Abstract

Effectively informing consumers is a big challenge for financial service providers. Triggering involvement in the personal situation of the client is a result of sending relevant information at the right time. While general machine learning techniques are able to accurately predict the behavior of consumers, they tend to lack interpretability. This is a problem since interpretation aims at producing the information a communication department requires to be able to trigger involvement. In this paper we provide a solution for predicting and explaining customer activation as result of a series of events, by means of deep learning and attention models. The proposed solution is applied to data concerning the activity of pension fund participants and compared to standard machine learning techniques on both accuracy and interpretability. We conclude that the attention based model is as accurate as top tier machine learning algorithms in predicting customer activation, while being able to extract the key events in the communication with a single customer. This results in the ability to help understand the needs of customers on a personal level and to construct an individual marketing strategy for each customer.

Keywords: Customer Activation, Deep Learning, Attention Models.

1 Introduction

For companies providing financial services, such as banks, insurers and pension funds, customer activation is crucial and often initiated in the best interest of their clients. While insight in their financial situation is of utmost importance to a client, they tend to postpone financial planning, due to the complexity and the lack of urgency (Lynch Jr and Zauberman, 2006). To create awareness, clients are often overloaded with information, this however results in choosing the path of least resistance and ignoring the information altogether (Agnew and Szykman, 2005). This is why understanding the needs of you clients is crucial. Assisting in their financial decisions comes down to communicating relevant information at the time it is needed. Therefore being able to recognize those moments (i.e. touchpoints) in life during which the client is in need of additional information, gives the company the opportunity to pro-actively provide the necessary answers. Clients searching for information, either as a result of a personal event or in response to communication sent by the company, are thus the ideal point of focus to extract these touchpoints.

In this paper we are therefore motivated, by application of machine learning, to predict if a client will pro-actively contact their financial service provider. Furthermore, we want to extract those events and/or characteristics of a client that led to the contact. The main focus of this paper lies therefore in the modeling that we propose concerning the problem of customer activation without a commercial motive. This includes both the predictive aspect and the interpretability of the results, as these should lead to actionable insights.

For this purpose, the paper makes use of data provided by APG, a large pension executioner in the Netherlands that deals with the pension plans of about 4.5 million people in the Netherlands. The data includes information on personal life events, demographical data and communications between the participant and the fund. The data will be structured as a sequence of events that either end up in contact with the fund, or not.

Given the sequential nature of the data used in this paper, it focuses on using models that can deal well with sequential data, such as deep learning (DL) and more specifically recurrent neural networks (RNN) (Hopfield, 1982). Deep learning is a subfield of machine learning concerned with the training of neural networks comprising many layers, which learn how to represent data at multiple abstraction levels (LeCun, Bengio, and Hinton, 2015; Schmidhuber, 2015). For the purpose of analysing sequential data, RNNs, networks in which the previous output of a unit is allowed to influence the next output, have shown promising results. As a further development, long-short term memory (LSTM) models (Hochreiter and Schmidhuber, 1997) improve and extend RNN networks, allowing for long sequences as input, thanks to the fact that they are less affected by the vanishing gradient problem (Hochreiter, 1998).

In addition to tackling the predictive (i.e. *'Will it happen?'*) part of the problem, there is the challenge of being able to interpret the results and answer the descriptive question (i.e. *'Why did it happen?'*). Attention models allow to focus the network on the part of the sequence that are of particular relevance to the output and therefore have the potential to help answer: *'Why will it happen in the future?'*.

This application of advanced neural networks to time dependent data of human behavior, while being able to interpret the (intermediate) results of the model, is the main contribution of this paper. For practical and ethical reasons, the possibility to interpret results while not losing predictive power has major implications in data driven decision making. In the modeling part of this paper, the LSTM network with attention is put into comparison, on both aspects, with standard machine learning models to deal with sequences of events.

The paper is organized as follows. Section 2 reviews previous research done on the subject of predicting customer behavior, as well as DL, RNNs and attention models in specific, and interpretability of machine learning models. Section 3 discusses the origin of the pension fund data and the way the sequence of events was extracted. Section 4 provides an in depth explanation of the proposed model. Section 5 consists of the results of the application of the model on the used data. Section 6 concludes with implications of the results and looks forward to possible future applications and improvements.

2 Related Work

Due to the technology driven approach to solve a practical business problem, prior research can be divided in two main subsections: the application (i.e. customer activation) and the technique (i.e. deep learning and attention models).

2.1 Customer activation

In relationship marketing and customer relationship management, customer engagement holds a vital position. Customer engagement is defined by Vivek, Beatty, and Morgan (2012) as "the intensity of an individual's participation in and connection with an organization's offerings and/or organizational activities, which either the customer or organization initiate". In this paper we focus on the latter part, the initiation that results in the engagement: customer activation. Customer engagement, and therefore customer activation, opens up the possibility for a company to construct a better relationship with the customer resulting in an increasingly more engaged and satisfied customer. Numerous research has shown the importance of customer satisfaction in relation to profitability (e.g. Anderson and Sullivan, 1993; Anderson, Fornell, and Lehmann, 1994; and Rust and Zahorik, 1993). For non-profit organizations or service providers with a fixed client base, customer satisfaction plays an even more crucial role. Due to the lack of a financial incentive, they fully focus on satisfying the needs of their customers. Hence being able to describe and analyse customer engagement behavior, potentially holds enormous value. According to Van Doorn et al. (2010) customer engagement behavior often results from motivational drivers. Exactly those drivers are crucial in understanding how to improve customer engagement. The fact that we collect increasing amounts of data on customers have made it possible to analyse customer behavior, this is what we call customer analytics. Customer analytics can be divided into three main areas (Nauck et al., 2006): customer segmentation, understanding customer views and predicting customer actions. We focus on predicting customer actions as we want to predict customer activation and understand the motivational drivers. A well-known example of this field of research is the prediction of churn or customer attrition. Here the correct identification of a potentially churning customer, gives the company the possibility to retain the customer. Over time classic machine learning techniques such as decision-tree (Wei and Chiu, 2002), support vector machines (Xia and Jin, 2008) and random forest algorithms (Xie et al., 2009), as well as neural networks (Sharma, Panigrahi, and Kumar, 2013) have been used with success to predict churning customers. While customer activation is not the same as churn, in many ways it is the complete opposite in fact, the prediction of these events are quite similar. Based on information of the customers such as demographics and prior behavior is predicted if they will show specific behavior.

2.2 Deep learning and attention models

Accurately predicting the behavior of a customer with specific characterizations, after certain life events or with a unique set of preferences, can quickly become very broad and complex. In theory, every small event could trigger a client to get actively involved in their financial situation, for example a simple discussion with a relative at a birthday party. This complexity and vast amount of (different) data sets, tend to be handled best by neural networks (Hopfield, 1982). A neural network is a set of multiple interconnected layers of neurons, inspired by the human brain, which can be trained to represent data at high levels of abstraction (LeCun, Bengio, and Hinton, 2015).

Neurons are artificial units, transmitting signals to the next neuron. Influenced by input from neurons in the previous layer or the input data, these neurons output a level of activity. By use of backpropagation, this network of neurons can be trained to recognize patterns and are able to predict the outcome of new data instances. Backpropagation is the gradient based learning method applied in the training of most neural networks.

Due to the differentiable activation functions, the network is able to back propagate the contribution of each neuron to the error of the training instance. This is used to update the weights of each of the neurons in order to achieve higher accuracy in the next iteration. Required is a known output matched to the input, therefore backpropagation is mostly used in supervised learning tasks. Backpropagation was first described by LeCun, Boser, et al. (1989) and is considered the accelerator neural networks needed to be further developed into an applicable algorithm. Deep learning is the collection of neural networks with multiple layers of neurons.

Given the sequential nature of the problem at hand (sequences of events), recurrent neural networks seem best suited. This specific type of a neural network architecture allow previous output of a unit to influence the next input and are therefore able to account for events in the past when predicting the next event. For example, RNN's are widely used in text translation, where the output (the translation) not only depends on a single word, but on the sequence of words.

General RNN's however suffer from the difficulty of learning dependencies over time. Due to the gradient based backpropagation algorithm, it gets harder to train the weights of the recurrent layers when the number of layers is increased (i.e. the length of the input sequence is increased), this is known as the vanishing gradient problem (Bengio, Simard, and Frasconi, 1994). Over time extended variations of the vanilla RNN models, Long-Short Term Memory (LSTM) and Gated Recurrent Units (GRU), became more popular due to the fact that those are less affected by this issue, allowing them to handle longer input sequences (Hochreiter, 1998).

LSTM's are a specific type of RNN architecture where information can be stored and removed from an internal memory cell (Hochreiter and Schmidhuber, 1997). These LSTM cell make use of four gates, in contrary to the single gate included in a general RNN cell, making them able to train what information to use from the input, what to forget and store in the memory cell and what to output in each state. Therefore being less susceptible to the vanishing gradient problem. In order to divide emphasis more equally over the input sequence and not specifically on the end, one can use a bidirectional LSTM (bi-LSTM). In this type of architecture there are two layers of hidden recurrent nodes, both connected to the input and output. However the second layer is differentiated by inputting sequences in a reverse order (Schuster and Paliwal, 1997).

This advanced RNN structure should be able to detect clients, with a specific set of characteristics or events, whom will be more or less likely to get actively involved in their pension. However, it is the particular event or set of events that caused the trigger in the life of the client that is most interesting and valuable to the business. This is why attention mechanisms will be researched and implemented on top of the deep learning architectures.

In a typical 'many-to-one' classifying problem tackled by a RNN, where a sequence is used to predict a single outcome, all information from the input sequence is summarized in the final hidden state of the recurrent layer. Attention models are able to not only use the intermediate hidden states, but can put more or less emphasis on previous hidden states. This technique has recently proved its value in visual recognition tasks (Xu et al., 2015), where the challenge was to describe the content of an image. The attention model was able to focus on the specific part of the image crucial in describing the next word of the output.

More recent developments in the field of attention models have either revolved around the application to more complex problems or combining the methodology with other advancements in the field of deep learning. Examples include visual question answering (Schwartz, Schwing, and Hazan, 2017), where the machine learns to answer a question based on (parts of) an image and Spatial Transformer Networks (Jaderberg, Simonyan, Zisserman, et al., 2015), which deals with the inability to be spatially invariant to (graphical) input data.

With the emphasis on accuracy with the latest developments in for example deep learning, models get more complex and therefore harder to interpret. This creates a tension between model performance and interpretabilty (Lundberg and Lee, 2017). However, interpretability of models is vital in understanding the results as well as in gaining the user's thrust in the results (Ribeiro, Singh, and Guestrin, 2016).

This is why a second advantage of applying attention models, next to the performance, is the explainability. Due to the weights trained by the model, the attention can be extracted and visualized, resulting in the possibility to pinpoint the part of the sequence or image that contributed most in each phase of the prediction (Xu et al., 2015). In our case, we should be able to apply attention to the sequence of (life)events in order to conclude which event(s) contributed most to the activation of the client.

3 Data

The data used in the construction and validation of the proposed model is provided by APG, a pension executioner in the Netherlands. In order to properly serve the participants of the pension funds it represents, the company needs information on the demographics of the population, e.g. gender, age and nationality. Furthermore, to accurately calculate the (future) entitled pension of a participant, it has data on the (alterations in the) life status of the participants, e.g. relationship status and employment history. In order to provide the best customer service, information concerning all contact the participant has with the company is stored and used for analysis, e.g. online (personal page and public website), phone and mail contact. Besides that, data regarding outbound communications is kept in order to continuously improve and optimize the communication with the participant.

The aforementioned data constructs the basis of the dataset used in this paper. The dataset will consist of three parts corresponding to participants on an individual level; the moment of activation, (the sequence of) other events, and demographics.

3.1 Definition of activation

Due to the objective of the proposed model, i.e. finding the trigger that results in activation, only participants are selected whom were inactive for at least two years (2014 and 2015) and shown interest in their pension afterwards (2016). Because of this, participants whom consistently check up on their pension, are left out. They would be considered to be already active by the fund. The inbound contact, which defines whether or not the participant gets active, is the visit to the personal online page of the participants. This is because the contact via this personal page is by far the most used way of retrieving information by the participant and the preferred channel of communication by the fund.

3.2 Sequence of events

In order to predict the behavior of a participant concerning the contact they have with their pension fund, a sequence of events is constructed that led up to the moment of contact. These events include 40 types of outbound communication (e.g. pension overviews and information concerning the fund, sent from the fund to the participant) and 11 different life events (e.g. marrying, divorcing, retiring and getting unemployed). Included are events that happened between 2014 and 2016 (but prior to the moment of contact). Padding was used to create sequences of consistent length over all the records.

3.3 Demographics

Next to the events that might influence the need for the participant to get actively involved in their pension, the demographical information is important to take into account. Next to some general information such as gender and age, included was information on possible triggers in the past to get involved in their pension, e.g. number of divorces and whether or not their partner is accruing pension at the same fund. The full list of variables and their corresponding values can be found in Table 1.

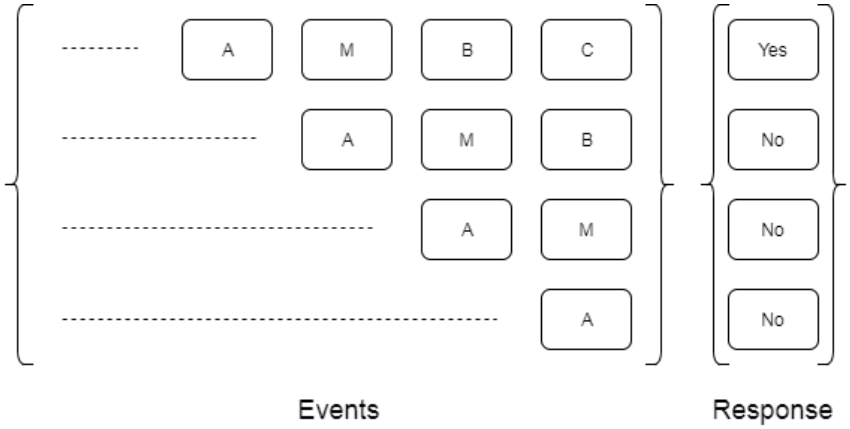
Variable	Type	Values
Gender	Binary	M/F
Age	Numerical	17 - 87
Marital Status	Categorical	Married/Never married/ Been married/Partnership
Residing in the Netherlands	Binary	Y/N
Retirement status	Categorical	Accruing/Idle/Retired
Preferred way of communication	Categorical	Paper/Digital/No preference
Newsletter subscription	Binary	Y/N
Number of divorces	Numerical	0 - 5
Partner at ABP	Binary	Y/N
Number of jobs	Numerical	0 - 32

Table 1: List of demographical information used with corresponding data type and the possible values.

3.4 Training data

The three aforementioned aspects make up the training examples. However these are all 'positive' instances, participants whom were activated after the series of events. To complete the training data, 'negative' instances are needed as well. These are constructed from parts of the sequence of events that did not (yet) led to contact with the participant. As an example: a participant consecutively received Campaign A, got married and received Campaigns B and C and subsequently logged in to his personal page (see Figure 1a). The training data then includes four sequences for this participant, three with no response and one with response (see Figure 1b).

(a) Example of a sequence of events.



(b) Transforming sequence of events into four records for training data.

Figure 1: Construction of training data.

Randomly 50,000 participants were selected out of those whom satisfied the requirements of being passive for at least two years before seeking contact with the fund. This resulted in 135,189 potential training sequences (an average of 2.7 sequences per participant and events leading up to activation). Table 2 displays the number of sequences that resulted in an activated participant versus the number of sequences that did result in activation.

Activated	Number of Sequences
Yes	50,000 (37.0%)
No	85,189 (63.0%)
Total	135,189 (100%)

Table 2: Distribution of positive and negative instances in the training data.

4 Methodology

The goal of the methodology we propose in this paper is to predict the possible activation of clients of financial service providers, and more specifically pension fund participants in this case, based on a sequence of events and characteristics of this person. This problem can be classified as a pattern recognition task and several machine learning techniques can be used to predict this behavior. Due to the sequential nature of the data we propose a deep learning algorithm: recurrent neural networks (RNN). To be even more specifically; the type of RNN used is a bidirectional long-short term memory (bi-LSTM) with an attention layer (see Figure 2).

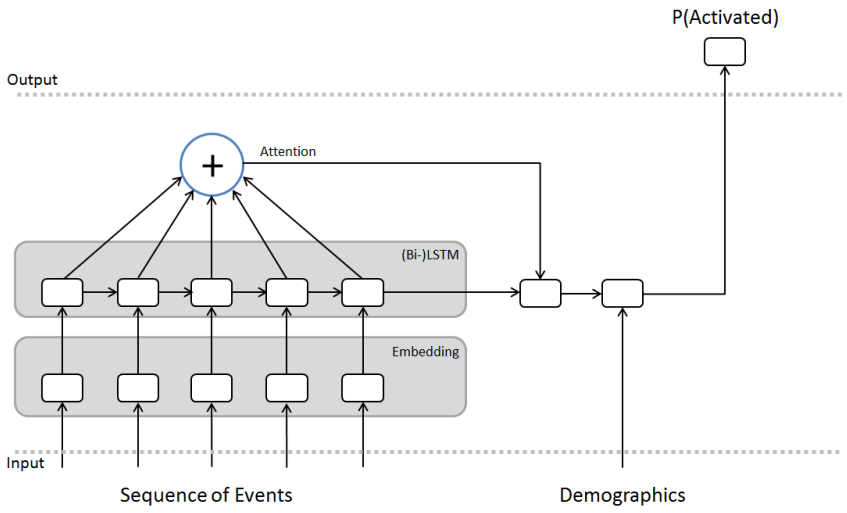


Figure 2: Model architecture.

The model has two different inputs and one output. The sequence of events, X_E, is the main input for the model. With n number of sequences as input and m events per sequence ($X_E \in \mathbb{N}^{n \times m}$). The demographical data, X_D is used as input in a later stage of the model. These features are one-hot-encoded and normalized in order to be fed into the model ($X_D \in \mathbb{R}^{n \times l}$). The output of this binary classification problem, $\hat{y}$, is a probability of the participant with a certain sequence of events and characteristics, to actively come in contact with the fund ($\hat{y} \in (0,1)^n$).

4.1 Embedding

Due to the fact that some events are related to others or often proceed a specific event, an embedding layer is used to plot the events in a higher dimensional space, with more similar events 'closer' to one another. The size of the embedding is a hyperparameter (i.e a model parameter that needs to be tuned). The output of the embedding layer is fed into a dropout layer (10%). Dropout is implemented to ensure a certain level of self-regularization during the training of the model. It ignores a part of the neurons (in this case 10%) during the forward or backwards pass in the training phase. This regularizes the importance of the neurons in a network, resulting in less overfitting in general.

4.2 Bi-LSTM

The output of the dropout is the input for the bi-directional LSTM. A regular LSTM layer consists of 4 gates, which control the input and output of the neurons. A LSTM-neuron is described by the following formulas (Tan, Xiang, and Zhou, 2015):

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f), \tag{1}$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i), \tag{2}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o), \tag{3}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ tanh(W_c x_t + U_c h_{t-1} + b_c), \tag{4}$$

$$h_t = o_t \circ tanh(c_t), \tag{5}$$

with b_f, b_i, b_o and b_c being biases. h_t is the hidden state of the neuron, which gets passed on to the next layer. It is the Hadamard product of the 'output-gate' o_t and the internal memory state c_t. The 'forget-gate' f_t and the 'input-gate' i_t regulate the balance between new input and previous information in the internal memory. This internal memory is what distinguishes an LSTM from a regular RNN, being able to store important information in the internal memory over a longer period.
The difference between a regular LSTM and the bi-directional LSTM used here is the order of the input from the sequences. The bi-LSTM consist of two independent LSTM layers, both with the same input sequences. However one layer goes over the sequence from start to finish and the other layer the other way around. This ensures the ability to pay equal focus on the whole sequence. The output of these layers are concatenated before fed into another dropout layer (10%).

4.3 Attention

The key section of the model however, is the attention layer, constructed on top of the bi-LSTM. Where the output of the LSTM is a hidden state (or in this case a concatenation of two hidden states), attention is able to focus on prior hidden states and combine (i.e. multiply by means of dot-product) this information with the output of the LSTM. This results in certain parts of the input sequence to have more or less impact on the outcome of the model. A trained attention layer is therefore able to recognize crucial information in an input sequence with regard to the output. This creates the benefit of interpretable output, due to the ability to extract the weights of this attention layer. In this case it yields information on which events in a sequence are more important in the prediction of activation. Attention can be formulated as follows (Bahdanau et al., 2016):

$$m_t = tanh(w_m h_t), \tag{6}$$

$$\alpha \propto softmax(w_a m_t), \tag{7}$$

with h_t the output of the LSTM at time t and w_m and w_a attention parameters. The attention and the output of the concatenated bi-LSTM are combined via dot product:

$$\tilde{h}_t = \alpha^T h_t. \tag{8}$$

After which another layer of dropout is added (10%).

4.4 Additional input

The output from the dropout is concatenated with the extra input from the demographical information corresponding to the participant going through the sequence of events. This concatenated output is passed on to three fully connected dense layers. The number of neurons in these layers, as well as prior layers, is a hyperparameter which needs to be tuned.

The last layer is a single output neuron with a sigmoid activation function that outputs a probability for the participant to login to their personal online page after a set of events.

4.5 Training the model

The goal is to predict inbound online contact from the client to the firm in the future. This is a binary classification problem, where 37.0% of the instances in the data are positive. In order to predict as accurately as possible, the challenge at hand becomes a minimization problem of the loss or error. In this case we minimize the binary cross entropy loss function:

$$L(\theta) = -\frac{1}{n}\sum_{i=1}^{n}(y_i log(\hat{y}_i) + (1-y_i) log(1-\hat{y}_i), \tag{9}$$

with θ the set of all trainable parameters, y_i the true label and $\hat{y}_i$ the predicted value of instance i.

5 Results

For reproducibility, we first discuss the hyperparameter tuning and the corresponding optimal values for the bi-LSTM model with attention described earlier, as well as for the models used as comparison later on. Hereafter we elaborate on the predictive performance of all models. Finally we will touch upon the interpretability of the tested models.

5.1 Grid search

In order to evaluate the model, the data is split into two parts: training set (80%) and test set (20%). This ensures that, while training the model with the training set, the evaluation can be performed on an out of sample data set.

For the hyperparameter tuning, or grid search, 10-fold cross validation is used on the training set. The parameters tuned in this process, with the respective ranges and the optimal value are shown in Table 3a.

Parameter	Range	Optimum
Number of neurons	[16,32,64,128]	**32**
Batch size	[32,64,128,256]	**128**
Number of epochs	[5:5:30]	**25**
Embedding size	[20,40,60,80]	**60**

(a) Bi-LSTM model

Model	Tuned parameters
k-NN	$k = 34$
GLM	$C = 1$
SVM	$kernel =$ 'RBF', $C = 10$, $\gamma = 1e-5$
XGBoost	$\eta = 0.2$, $max\_depth = 5$

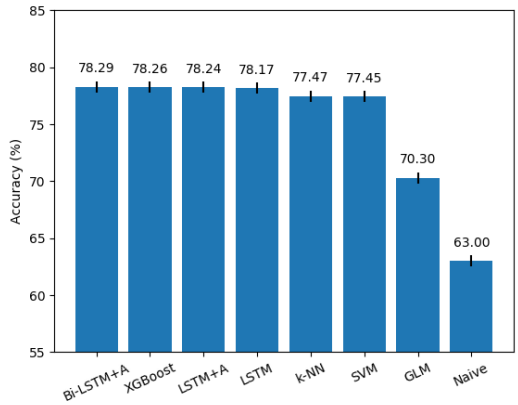
(b) Other models

Table 3: List of hyperparameters tuned during grid search, with their respected optimal values.

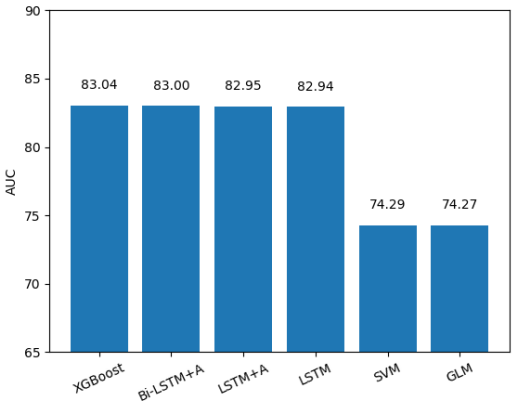
In order to assess the quality of the model, its results will be compared to the performance of other, non neural network, algorithms. These include k-nearest neighbor (k-NN), generalized linear model (GLM), support vector machine (SVM) and extreme gradient boosting (XGBoost). The models with tuned parameters are shown in Table 3b.

5.2 Performance

The performance is evaluated by means of the accuracy and the area under the ROC-curve. The proposed bi-LSTM with attention is able to accurately predict the outcome of 78.29% of the input sequences in the test set and the area under the ROC-curve is 83.00. Figures 3a and 3b depict the performance of each of the different models for comparison. As one can see, the performance of the bi-LSTM model with attention is very similar to three of the other models. With $CI_{95\%} = [77.79, 78.78]$ for the accuracy of the bi-LSTM with attention, none of the results of the three other models out of the top four is significantly different. Hence we can conclude that the attention has little to no impact on the performance of the LSTM models. As well as the fact that the bi-directionality of the main model did not improve the performance of the model. The XGBoost model performs on par with the LSTM architectures, which was unexpected.



(a) Accuracy (with 95% confidence interval).



(b) Area under the curve (AUC).

Figure 3: Performance of the models.

5.3 Interpretability

Besides the performance of the proposed model, the interpretability is crucial in applying the results in practice. The attention layer gives us the possibility to extract the importance of each event in a single

input sequence. An example is given in Figure 4: it shows an input sequence of three events (and seven padded non-events) from least recent to most recent. According to the weights, we can conclude that starting a new job is an important event in the life of the participant, regarding his or her activity towards pension. The two (different) offline campaigns have very little influence on this behavior. Furthermore we see that the padded events hold information as well. Apparently the fact that there is only three events in this sequence increases the possibility of activation. This could be due to customers being less and less probable to be activated every time they not respond to outbound communication. Hence the relatively high focus on the non-events preceding the actual events.

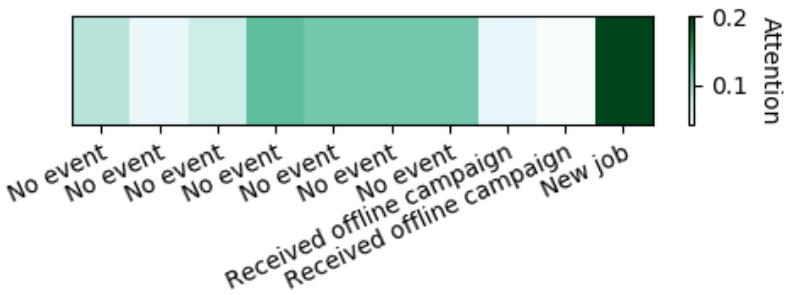


Figure 4: Example of attention divided over a single sequence leading up to activation.

The possibility of extracting the crucial events for a specific client, is what distinguishes this technique from other semi-interpretable machine learning techniques. For example XGBoost is able to provide the user with aggregated variable importance based on information gain (see Table 4). This however is not applicable to a single instance and therefore less insightful in this case, where the pension fund strives for personal and individual contact.

Type	Content	Information gain
Offline magazine	Status update fund (Nov15)	0.203
Offline magazine	Status update fund (Feb16)	0.075
Offline magazine	Status update fund (Mar17)	0.064
Life event	New job	0.048
Mail	Status update fund (Mar17)	0.046

Table 4: Top five events with highest information gain as results of the XGBoost model.

However, when we do take a closer look at the average activity for each of the events, we see a clear set of events that have a greater effect on activating the participants (Table 5). Apparently a very specific campaign in 2014, where call agents pro-actively called participants welcoming them to the fund after starting a new job, spiked the activity of those participants. Furthermore, campaigns revolving losing a job, switching jobs, working part-time and retiring part-time are quite effective. These events can be seen as ‘touchpoints’ in the lives of the participants where they are more interested in their pension, compared to other events such as getting married or divorcing. With this information, pension funds are able to target their participants with more relevant information at the right time in their lives.

Type	Content	Number of occurrences	Average activity
Outbound call	New fund participant	199	0.198
Letter	Parttime retirement	4853	0.169
Life event	Losing job	3597	0.158
Mail	Celebrating 50^{th} birthday	1069	0.148
Mail	Parttime retirement	1576	0.147

Table 5: Top five events with overall highest activity as result of the attention model.

6 Conclusion

This paper presented a modeling of the problem of customer activation for financial service providers. Data concerning the customer interaction, life events and their demographics were used in order to model each of the customers as complex sequences of events. Deep learning models with attention were used in order to obtain a numerical explanation concerning the events that had more effect on activating a single customer.

We devised a deep learning solution to predict customer activity in a non commercial setting. The goal was two-fold: predicting customer activation by applying deep neural networks to the sequence of events that may lead to this activation concerning their financial situation and extracting the crucial event(s) that motivated this activity. This to provide actionable insights to financial service providers, concerning which of their clients most likely is in need of more information and can be (pro-actively) assisted with his financial questions.

Data was used of $50,000$ Dutch pension plan participants showing pension related activity after inactivity for a period of at least two years. Included was information on life events, demographics and campaigns sent from the fund to the participant.

The proposed model is an adaptation of the classic recurrent neural network; a bidirectional long-short term memory model with a layer of attention. In order to evaluate the performance of the model, it was compared to standard machine learning algorithms and neural networks without attention. With an accuracy of 78.29% of correctly predicting if a sequence of events results into activation, the model performed similar to a widely used machine learning solution (i.e. Extreme Gradient Boosting) and LSTM models without attention.

However the main contribution of this research lies in the interpretability of the model with attention and the resulting insights. As most machine learning solutions provide a prediction without explanation in the process or an aggregated summary, the attention based model was able to extract the events that more often than not result in contact with the company, even on an individual level.

6.1 Discussion

While evaluating the trained models we expected the bi-LSTM to perform better in terms of predictive power (accuracy and AUC) than the less complex algorithms, due to the complexity of the data. Even though there was a significant higher performance when compared to k-NN, SVM and GLM, the proposed solution did not significantly outperform the more basic LSTMs or XGBoost. In hindsight, we believe this to be due to the relatively simple model objective, that is the binary classification. As of now we defined customer activation specifically as a login to one's personal page, but this could be extended with other possible signs of activation (e.g. calling, mailing or chatting with the fund). This would yield further insight in the effect of events and campaigns on the type of activation, which can be used to direct customers in the future to the communication channel preferred by the customer or company.

Even with similar performance, the found results are still promising, as companies would be able to act on the insights, while being convinced by the predictive performance of the model. For instance,

companies would be able to personalise their contact with the customers, while only communicating relevant information to clients in need of it, i.e. sending the right information to the right customer at the right time. This would help decrease the information overload, discussed by Agnew and Szykman (2005), that often results in ignoring the information completely. Therefore the information that is communicated, has a larger impact.
If we take the example of Figure 4 for instance, the marketing department could decide to send information regarding the impact of starting a new job on your retirement plan to this specific individual. Moreover they can assess the impact of sending a campaign to a participant, as they would be able to, in advance, compare the impact of different campaigns after an event. Figure 5 shows the difference in probability of activation in case either a card gets sent or the participant is welcomed with a call after starting a new job. This gives them the opportunity to perform a cost-benefit analysis and decide upon the preferred way of communication. This creates the possibility to construct a full marketing strategy to activate a single participant, based on their personal situation.
Besides the practical implications of being able to explain the results, there is an ethical benefit as well. Due to the fact that the importance of predictive performance of a model has hugely increased, interpretability has moved to the background (Lundberg and Lee, 2017). Applying these so called 'black box' models on human behavior can (often unwillingly) create a bias towards specific groups of individuals. Being able to explain *why* the model prescribes sending information to an individual creates the possibility to check for these biases.

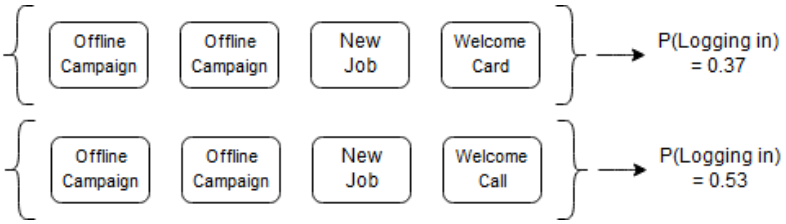


Figure 5: Difference in probability of activation between two marketing strategies responding to a participant with a new job.

6.2 Future work

The results create opportunities for future research as well. First of all, the model can be empirically tested and used to increase the effectivity of communication with the clients. As explained, an optimal chain of campaigns can be devised, responding to events marked as important in relation to activating the participant. This should potentially lead to higher customer satisfaction and money saved on ineffective communication.
Secondly, LSTM models are not the only neural network architectures that handle sequential data well. Gated recurrent unit (GRU) networks and convolutional neural networks should be tested on the same data set to assess the quality of the architectures. These can also be extended with attention models.
Thirdly, as we see that the model performs similar to other machine learning solutions, the algorithms could be exposed to a more detailed set of information. This is a larger set of participants and events, and more detailed information on the communication with the participant. Besides that, the objective can be adjusted to not only predicting activation, but the way the participant will contact the firm as well. This is where neural networks tend to outperform standard solution as the amount of information and the complexity of the problem increases.
Finally, the fact that we were able to pinpoint those (type of) events that result in contact from the participant to the fund, might shift the objective of the model towards the prediction of those specific events. This calls for a new setup and data set and interacts with the possibility of acquiring new sources of information (e.g. external data).

References

Agnew, J. R. and L. R. Szykman (2005). "Asset allocation and information overload: The influence of information display, asset choice, and investor experience." *The Journal of Behavioral Finance* 6 (2), 57–70.

Anderson, E. W., C. Fornell, and D. R. Lehmann (1994). "Customer satisfaction, market share, and profitability: Findings from Sweden." *Journal of marketing* 58 (3), 53–66.

Anderson, E. W. and M. W. Sullivan (1993). "The antecedents and consequences of customer satisfaction for firms." *Marketing science* 12 (2), 125–143.

Bahdanau, D., J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio (2016). "End-to-end attention-based large vocabulary speech recognition." In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*. IEEE, pp. 4945–4949. ISBN: 978-1-4799-9988-0. DOI: 10.1109/ICASSP.2016.7472618. URL: https://doi.org/10.1109/ICASSP.2016.7472618.

Bengio, Y., P. Simard, and P. Frasconi (1994). "Learning long-term dependencies with gradient descent is difficult." *IEEE transactions on neural networks* 5 (2), 157–166.

Hochreiter, S. (1998). "The vanishing gradient problem during learning recurrent neural nets and problem solutions." *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 6 (02), 107–116.

Hochreiter, S. and J. Schmidhuber (1997). "Long short-term memory." *Neural computation* 9 (8), 1735–1780.

Hopfield, J. J. (1982). "Neural networks and physical systems with emergent collective computational abilities." *Proceedings of the national academy of sciences* 79 (8), 2554–2558.

Jaderberg, M., K. Simonyan, A. Zisserman, et al. (2015). "Spatial transformer networks." In: *Advances in neural information processing systems*, pp. 2017–2025.

LeCun, Y., B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel (1989). "Backpropagation applied to handwritten zip code recognition." *Neural computation* 1 (4), 541–551.

LeCun, Y., Y. Bengio, and G. Hinton (2015). "Deep learning." *Nature* 521 (7553), 436–444.

Lundberg, S. M. and S.-I. Lee (2017). "A unified approach to interpreting model predictions." In: *Advances in Neural Information Processing Systems*, pp. 4765–4774.

Lynch Jr, J. G. and G. Zauberman (2006). "When do you want it? Time, decisions, and public policy." *Journal of Public Policy & Marketing* 25 (1), 67–78.

Nauck, D., D. Ruta, M. Spott, and B. Azvine (2006). "Being proactive—analytics for predicting customer actions." *BT Technology Journal* 24 (1), 17–26.

Ribeiro, M. T., S. Singh, and C. Guestrin (2016). "Why should i trust you?: Explaining the predictions of any classifier." In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, pp. 1135–1144.

Rust, R. T. and A. J. Zahorik (1993). "Customer satisfaction, customer retention, and market share." *Journal of retailing* 69 (2), 193–215.

Schmidhuber, J. (2015). "Deep learning in neural networks: An overview." *Neural networks* 61, 85–117.

Schuster, M. and K. K. Paliwal (1997). "Bidirectional recurrent neural networks." *IEEE Transactions on Signal Processing* 45 (11), 2673–2681.

Schwartz, I., A. Schwing, and T. Hazan (2017). "High-Order Attention Models for Visual Question Answering." In: *Advances in Neural Information Processing Systems*, pp. 3664–3674.

Sharma, A., D. Panigrahi, and P. Kumar (2013). "A neural network based approach for predicting customer churn in cellular network services." *arXiv preprint arXiv:1309.3945*.

Tan, M., B. Xiang, and B. Zhou (2015). "LSTM-based Deep Learning Models for non-factoid answer selection." *CoRR* abs/1511.04108. URL: http://arxiv.org/abs/1511.04108.

Van Doorn, J., K. N. Lemon, V. Mittal, S. Nass, D. Pick, P. Pirner, and P. C. Verhoef (2010). "Customer engagement behavior: theoretical foundations and research directions." *Journal of service research* 13 (3), 253–266.

Vivek, S. D., S. E. Beatty, and R. M. Morgan (2012). "Customer engagement: Exploring customer relationships beyond purchase." *Journal of marketing theory and practice* 20 (2), 122–146.

Wei, C.-P. and I.-T. Chiu (2002). "Turning telecommunications call details to churn prediction: a data mining approach." *Expert systems with applications* 23 (2), 103–112.

Xia, G.-e. and W.-d. Jin (2008). "Model of customer churn prediction on support vector machine." *Systems Engineering-Theory & Practice* 28 (1), 71–77.

Xie, Y., X. Li, E. Ngai, and W. Ying (2009). "Customer churn prediction using improved balanced random forests." *Expert Systems with Applications* 36 (3), 5445–5449.

Xu, K., J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio (2015). "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention." *CoRR* abs/1502.03044. URL: `http://arxiv.org/abs/1502.03044`.