Article **Journal Information Technology**
December, 2023 Vol.10 No.30 27-33

Recommendation of account effect by CFDI Based on Machine Learning

Recomendación de cuentas afectadas por CFDI basado en aprendizaje automático

MONTECILLO-PUENTE, Francisco Javier & PERÉZ-MONCADA, Santiago

Tecnológico Nacional de México campus Salvatierra (ITESS)

ID 1st Author: *Francisco Javier, Montecillo-Puente* / **ORC ID**: 0000-0001-9540-9228, **Researcher ID Thomson**: X-2309-2018, **CVU CONAHCYT ID**: 50009

ID 1st Co-author: *Santiago, Peréz-Moncada*

DOI: 10.35429/JIT.2023.30.10.27.33 Received: August 12, 2023; Accepted December 20, 2023

Abstract

This work presents a neural network model for recommending accounts that a CFDI affects. FactureApp is an invoicing application with different services developed by the company LionDev S.A de C.V. The accounting module for generating policies based on Boolean algebra is one of their main services. This application implements the laws established in the Federal Tax Code and applied by the Tax Administration Service SAT of Mexico. This work defines the elements of the CFDI that can be used to determine the account it will affect. Also, the modeling and training of neural network with 327 input data, a hidden layer with 364 neurons and an output layer of size 200 is presented. Python and the Keras module of TensorFlow are used to model and train the network. Also, an API based on the Flask Framework is modeled for requesting recommendations. At the end of this article, improvements in the model and work that must be developed to integrate it into production are discussed.

FactureApp, machine learning, SAT and CFDI

Resumen

En este trabajo se presenta el modelo de una red neuronal para la recomendación de cuentas que un CFDI afecta. FactureApp es una aplicación de facturación con diferentes servicios desarrollada por la empresa LionDev S.A de C.V. Uno de estos es la modulo contable de generación de pólizas basadas en algebra booleana. Esta aplicación implementa las leyes establecidas en el Código Fiscal de la Federación y aplicadas por el Servicio de Administración Tributaria SAT de México. En este trabajo se definen los elementos del CFDI que pueden ser usados para determinar las cuenta de este va a afectar. También, se presenta el modelado y entrenamiento de una red neurona con 327 datos de entrada, una capa oculta con 364 neuronas y una capa de salida de tamaño 200. Para el modelado de la red se utiliza Python y el módulo Keras de TensorFlow. También, se modela una API basada en el Framework flask para la solicitud de recomendaciones. Al final del presente artículo se discuten mejoras en el modelo y trabajo que se debe desarrollar para integrarlo a producción con la aplicación.

FactureApp, aprendizaje automático, SAT y CFDI.

Citation: MONTECILLO-PUENTE, Francisco Javier & PERÉZ-MONCADA, Santiago. Recommendation of account effect by CFDI Based on Machine Learning. Journal Information Technology. 2023. 10-30: 27-33

*Author's Correspondence: (e-mail: famontecillo@itess.edu.mx)
†Researcher contributing as first author.

1. Introduction

One of the most widely used tools for problem solving is Machine Learning, (Sarker, 2021). ML has been used in image classification systems, speech recognition, text generation, virtual actor animation, automatic robot navigation, to name a few (Negi and Rajesh, 2019). In terms of invoice analysis, clafication systems have been developed, a task that was previously performed manually (Tarawneh et. al, 2019). In (Bardelli, *et al*, 2020) a paper is presented where the electronic invoice in XML format is examined to facilitate some accounting tasks such as determining the nature of bank transfers.

In particular, this work presents a neural network to simplify the process of creating accounting policies. This process consists in defining the conditions to generate the digital tax receipts CFDI's (SAT, 2023) stamped in the FactureApp platform (Facture, 2023), then some rules are created to define the accounting accounts that will be affected. Finally, the policy is generated with its respective accounting entries. In this work a neural network is trained to define a policy recommendation system based on the possible accounting accounts that can modify the CFDIs. This is achieved by creating a neural network model and training it with CFDI's and policies already generated.

The Facture App platform of LionsDev S.A. de C.V., in order to automate the creation of accounting policies, requires the end user, who is generally an accountant, to enter a set of rules. These rules are constructed using Bool algebra. Here, it challenges users to train and develop Boolean algebra skills. Python (Python, 2023) and the modules: tesnsorflow (Abadi *et al*., 2015), pandas, numpy, keras, sklearn and flask were used to implement the neural network. It is used in the Anaconda development environment.

2. General aspects of machine learning and fundamentals of accounting issues

In this section the fundamental concepts of machine learning and policy generation are presented. As well as, the definition of the data of interest.

2.1. General steps for problem solving using machine learning

Machine learning follows the following phases for problem solving: data generation, data pre-processing, model generation, model evaluation, parameter tuning and release.

2.2. Accounting policies

An accounting policy is a physical or digital document that reflects the accounting movements of an organisation. It provides an accurate record of its operations and helps to make financial decisions. An accounting entry is the notation or record that reflects some economic movement of a company or person. This movement can be an inflow or outflow of money. In general, accounting entries are assets, liabilities and equity.

An accounting policy must contain the following data: a) information on the type of policy; b) item with accounting account, sub-account and auxiliary accounts; c) CFDI vouchers; d) taxes, e) RFC and the amount covered by the policy. There are different types of policies. A journal policy records and attaches the vouchers of business transactions that do not generate bank movement in the business account. The outgoing policy records all outgoing cash movements, e.g. payroll payments, payments to creditors, among others. A third type of policy is the income policy, which records movements that generate an inflow of money to the company's account.

In Mexico, Article 28 section. IV of the Fiscal Code of the Federation (CFF) mentions that the persons who, in accordance with the fiscal provisions, keep the accounting of the company must comply with the monthly entry of the accounting information, through the official page of the SAT, in the rule 2.8.1.6 of the SAT talks about the compliance of the delivery of the accounting in electronic media in a manual way. In general terms, it mentions that the information submitted must be in XML format; legal entities must send the accounting information no later than the first 3 days of the second subsequent month; it mentions that individuals must send it monthly no later than the first 5 days of the second subsequent month; it also states that issuing taxpayers must send the information in monthly files every four-month period.

ISSN: 2410-4000

This work is done using the Facture App platform. Within the modules, there is the accounting automation module. This tool is used to reduce and optimise the work time of the accountants. It creates rules and conditions for the initial configuration of a chart of accounts where the policies of the accounting entries will be created automatically.

To automate the creation of policies, it is first necessary to define the condition that allows to obtain the CFDIs that are of interest, for example, PPD (payment in deferred instalments) that belong to an issuer with a particular RFC. At this point, the rules that must be complied with are defined for each receipt or concept. The rules that are defined have the form SI (condition1) OP (condition2) OP (...), where OP refers to a Boolean operator.

An example of a rule is:

IF (RFC of issuer contains IVDNNNNNN)
and
(Voucher payment method EQUALS PPD)

Once the rules have been created, they must be associated to the accounting records. Subsequently, the rule is applied to identify the CFDIs that comply with the rules and from these the policies are generated automatically.

Within the company Liondev SA de CV it was found that the end users, accountants, have difficulties in the use of the automation module for the generation of policies. Specifically in the definition of the rules to filter the CFDIs of interest for the creation of accounting policies. This paper addresses the use of IA for the creation of accounting policies from a CFDI as a source document.

2.3. Definition of the dataset variables

Based on the previous section, the attributes of interest of a CFDI XML are payment method, voucher type, payment method, CFDI usage and product or service key. A CFDI receipt is shown in Figure 1.

Figure 1 XML of a CDFI, with its different fields

The values of the payment method variable are: PUE (payment in a single exhibition) and PPD (payment in instalments or deferred). For the voucher type variable: I (income), E (expenditure), T (transfer), N (payroll) and P (payment) are used. The form of payment is a variable that indicates the means by which the products or services can be paid, see Table 1.

Method of payment	Description
01	Cash
02	Nominative cheque
03	Electronic funds transfer
04	Credit card
05	Electronic purse
06	Electronic money
08	Food vouchers
12	Payment in kind
13	Payment by subrogation
14	Payment by consignment
15	Forgiveness
17	Compensation
23	Novation
24	Confusion
25	Remission of debt
26	Prescription or lapse of time
27	To the satisfaction of the creditor
28	Debit card
29	Service card
30	Advance payment application
31	Payment intermediary
99	To be defined

Table 1 Values for the form of payment variable.

The CFDI use variable is the option that best describes the use of the CFDI for deductions. Given that the list is very long, only some of these are shown: G01 (Acquisition of goods), G02 (Returns, discounts or rebates), G03 (General expenses), I01 (Construction), D01 (Medical fees), S01 (No tax effects), CP01 (Payments), CN02 (Payroll) and P01 (To be defined). For the product or service code, this allows you to identify the concept you wish to invoice, in the catalogue provided by the SAT it has 52 thousand options.

ISSN: 2410-4000

This code is made up of 8 digits: division (first two digits), group (next two digits), class (next two digits), subclass (last two digits). For example, the code 50201706 refers to the following:

– 50: Food, beverages and tobacco.
– 20: Beverages
– 17: Coffee and tea
– 06: Ground coffee.

In order to use this type of information, it must be converted to numerical values. One technique is using cardinality of the dataset, another technique is applying hash code (when the amount of data is very large).

3. Machine learning model for generating account recommendations

The model proposed is a neural network for data classification. Given the input sets, the response of the network is a probability vector, where each value indicates the probability of belonging to a class. In particular, the resulting vector in our case will indicate the probability of making movements in the accounting accounts.

3.1. Input data

To train the network, an initial dataset of dimension 5 with 2226 records was created. This data has the format shown in Table 2.

No	Method of payment	Method of payment	Type of voucher	Use of CFDI	Product or service key
0	PUE	1	INGRESO	I02	01010101
1	PPD	99	INGRESO	S01	10101500
2	PUE	2	INGRESO	I05	10101500
2225	PUE	99	NOMINA	I04	10224700

Table 2 Format of the data record. The product or service key was separated according to the form in which it is coded. Therefore the record has dimension 8

The data set consigns 200 accounting accounts represented by consecutive numbers. For the above input data, their output values are shown in Table 3.

No.	y(exit
0	**0**
1	**1**
2	**2**
2225	**37**

Table 3 Table of desired output values, of the accounting accounts to be assigned

Since the data to be trained are few, this dataset was increased. Using two strategies: randomly duplicating a record and artificially generating new ones with known data up to 71232.

Now the coding of the data was done using the One-Hot technique (Hancock, 2020), where all possible values for each input variable are presented and a value of 1 is assigned only to the value corresponding to the value of the input in Table 2. This results in an input vector of 327 binary entries, thus obtaining a sparse matrix of 71232 rows x 327 columns, the rows corresponding to the number of records.

This data set is divided into two sets, a test set and a training set, divided into 25% and 75% respectively.

3.2. Definition of the neural network model

To define the neural network, the TensorFlow keras library is used. The network is composed of an input layer of 327 neurons; a hidden layer of 364 neurons with ReLu activation function and L2 regulator; and finally an output layer with 200 neurons with softmax activation function, see Figure 2.

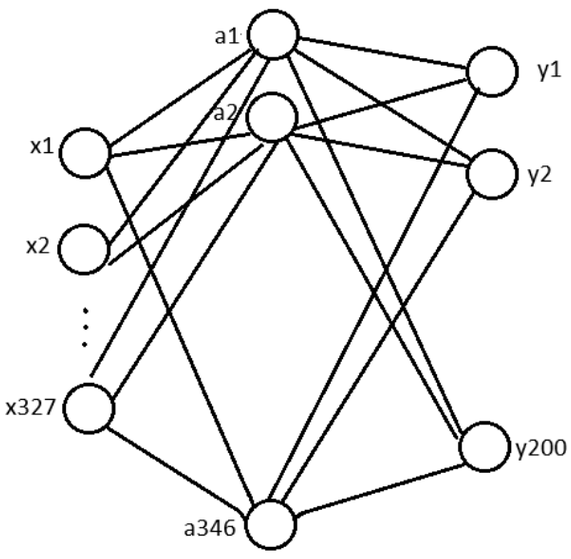
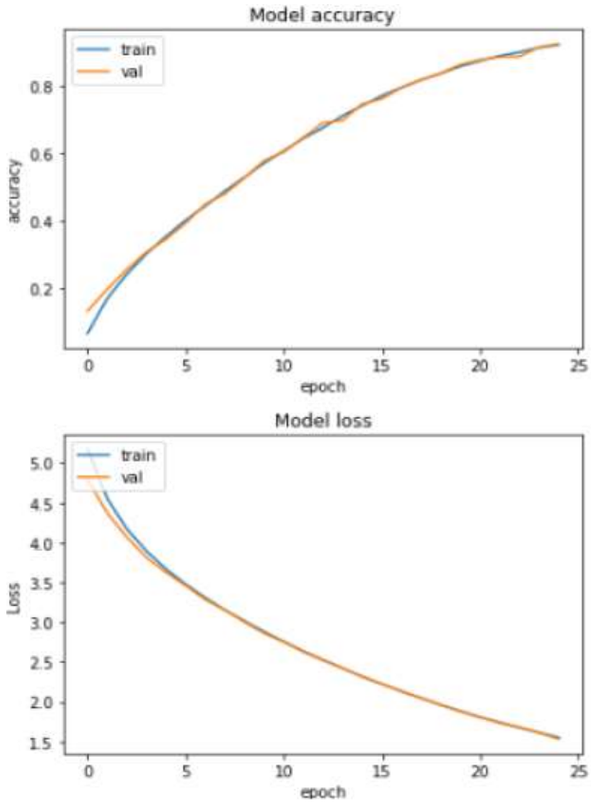


Figure 2 Neural network with 327 inputs, a hidden layer of 346 neurons and 200 outputs

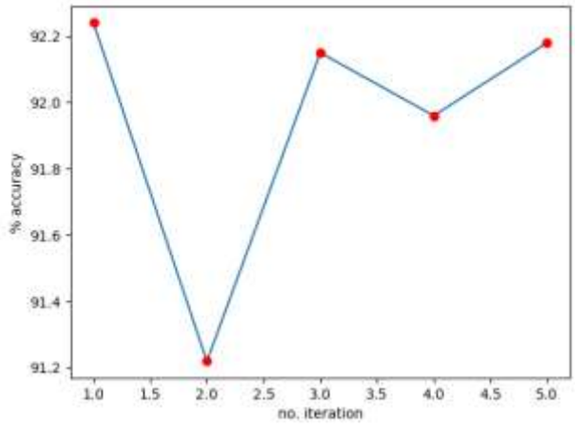
3.3. Training the neural network

To train the neural network, the cross-validation technique defined in 5 iterations was used. The validation consists of evaluating by training using subsets of data and evaluating with the complementary set.

It was defined 75% for training and 25 for validation. In addition, 25 epochs per iteration are used. Figure 1 shows the loss function and the percentage accuracy for the first iteration of training. In Graph 2, the evolution of the accuracy over the 5 iterations is presented. From the graph an accuracy of 92.18% is obtained, hence in this model in the first iteration a maximum value of 92.24% is obtained. However, the use of cross-validation and training by epochs is intended to avoid over-fitting the network over-fitting of the network.



Graph 1 Model accuracy and network loss function, during first iteration.



.

Graph 2 Accuracy of the model during the 5 iterations, it is observed that the maximum value is found in the first iteration.

The trained model is stored in a h5 and json format for future use. The JSON file stores all the characteristics and architecture configurations of the neural network and the .h5 file stores the value of the neural network hyperparameters generated in the training process.

4. API for prediction of accounting accounts that modifies the CFDI

To create the API, tensorflow was used to load and evaluate the model; sklearn for One-Hot encoder, numpy matrix operations and the mini framework flask for creating web applications. The prediction API workflow consists of first loading the prediction model. The json format consists of the format shown in Figure 3. Within the format are the fields from section 3.1, as well as the number of recommendations that correspond to the accounts that this CDFI should modify. With the data included in the request, the input values are created using sklearn's One-Hot.

```
{
    "data":{
        "metodoDePago": "PUE",
        "formaDePago": "1",
        "tipoDeComprobante":"INGRESO",
        "usoCFDI":"I02",
        "claveProdServicio": "01010101"
    },
    "numberRecommendations":5
}
```

Figure 3 json representation of the request

Subsequently, the input is submitted and the model is evaluated. The response is generated by sorting the results of the network, the first 5 are encoded and sent in a json format. The format of a response is shown in Figure 4.

```
{
    "prediction": [
        1,
        191,
        135,
        143,
        87
    ],
    "text": "Esta es la predicción..."
}
```

Figure 4 Request response format, indicating the possible categories that the CFDI is to be modified

5. Results obtained

A first result was the analysis of the CFDIs to determine the structure of the input and output data using the One-Hot technique. At this point, a proprietary database was generated in which, given a CFDI, the account to be modified is categorised. A second result obtained was the creation of a neural network for the suggestion of the categories or accounting accounts that are likely to modify the CFDI. During training, an accuracy of 92.18% was obtained. Also, a web API was created to make recommendations of the movements of accounting accounts that are modified by a CFDI.

6. Acknowledgement

We would like to thank the Tecnológico Nacional de México campus Salvatierra for the development of this work, as well as the company LionDev S.A de C.V. for providing sample data generated by its Facture App application.

7. Conclusions

This work presented a neural network model for account recommendations that a CFDI can modify. Fields of a CFDI that are of interest to determine the account were defined, and how to code these entries was also presented. However, further collection of information from records needs to be analysed to train the model so that it can be moved into production. Also, the creation of different models with more neurons and more hidden layers can be explored. More experimentation with real data and end users is needed. Another area of opportunity is the development of interfaces that allow users to query and collect data to train the network or modify the model.

References

Bardelli, Chiara, Alessandro Rondinelli, Ruggero Vecchio, and Silvia Figini. (2020). Automatic Electronic Invoice Classification Using Machine Learning Models Machine Learning and Knowledge Extraction 2, no. 4: 617-629. https://doi.org/10.3390/make2040033

Facture. (2023) https://facture.com.mx/

Flask. (2023) https://flask.palletsprojects.com/en/3.0.x/

Hancock, J.T., Khoshgoftaar, T.M. (2020) Survey on categorical data for neural networks. J Big Data 7, 28. https://doi.org/10.1186/s40537-020-00305-w

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo,Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Rafal Jozefowicz, Yangqing Jia, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Mike Schuster, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker,Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. (2015) TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Negi A. and Rajesh, K. (2019). A Review of AI and ML Applications for Computing Systems," 2019 9th International Conference on Emerging Trends in Engineering and Technology - Signal and Information Processing (ICETET-SIP-19), Nagpur, India, 2019, pp. 1-6, doi: 10.1109/ICETET-SIP-1946815.2019.9092299.

Numpy. (2023). https://numpy.org/

Python. (2023) https://www.python.org/

Pandas. (2023). https://pandas.pydata.org/

SAT. (2023) https://www.sat.gob.mx/aplicacion/75169/servicio-de-facturacion-cfdi-version-4.0-(vigente-a-partir-del-1-de-enero-de-2022)

Scikit-learn. (2023) https://scikit-learn.org/stable/ gob.mx. (2023) https://www.gob.mx/indesol/documentos/codigo-fiscal-de-la-federacion-64540

Sarker, I.H. (2021) Machine Learning: Algorithms, Real-World Applications and Research Directions. SN COMPUTER. SCIENCE. 2, 160 (2021). https://doi.org/10.1007/s42979-021-00592-x

sat. (2023) .https://www.sat.gob.mx/home

sat-catalogo. (2023) https://www.sat.gob.mx/consultas/53693/catalogo-de-productos-y-servicios

Tarawneh, Ahmad & Hassanat, Ahmad & Chetverikov, Dmitry & Lendak, Imre & Verma, Chaman. (2019). Invoice Classification Using Deep Features and Machine Learning Techniques. 10.1109/JEEIT.2019.8717504.