Data Science Report

Application Exercise, Konfío

Andrés Troiano

July 5, 2021

1 Business Understanding

Konfío is a Mexican fintech company that provides funding for small and medium-sized businesses. During the evaluation process for potential clients, applicants are asked to provide a short written use of proceeds. That is, a description of how they are planning to spend the capital. Understanding these intentions is very important for Konfío, so the company would like to be able to classify these descriptions in different categories. Some examples are: to pay debt, to buy new equipment, or for payroll.

The company has tested two approaches to solve this problem: Machine Learning (SVC, LogReg, RandomForest, Multinomial NBayes) and Deep Learning (CNN). The best performing algorithm was Logistic Regression, with a loss score of 0.068 and an average precision of 0.67. The goal is to provide an alternative solution that hopefully performs better.

2 Data Understanding

The dataset provided is made of 11 columns. The first one contains the uses of proceeds. That is, each row has the description of the intentions of one applicant. The remaining columns contain the provided labels for the texts. The categories are as follows:

- crec: insinuation of a strategy for growth, even if no specifics were provided.
- cred: debt payments.
- equ: purchasing equipment.
- inic: kickstarting the business.
- inv: purchasing inventory.

- mkt: marketing.
- no: not destined to working capital. In practice this often means personal or private expenses, like medical treatments or schooling.
- renta: pay the rent.
- sueldo: payroll.
- temp: seasonal expenses.

The dataset has one column for each of these categories, with a value of 1 every time the description falls into said category, and 0 every other time. More than one labels are allowed. The dataset totals 6679 uses of proceeds. The first five rows are shown in figure 1:

	motivos	crec	cred	equ	inic	inv	mkt	no	renta	sueldo	temp
0	Crear un departamento de ventas e inversión a ...	0	0	0	0	0	1	0	0	0	0.0
1	establecerme en un local y agregar materia pri...	0	0	0	0	1	0	0	1	0	0.0
2	Compra de equipo e incrementar inventario	0	0	1	0	1	0	0	0	0	0.0
3	Invertir en crecimiento de flotilla de unidade...	0	0	1	0	0	0	0	0	0	0.0
4	Para comprar mercancía y comprar lonas nuevas	0	0	0	0	1	0	0	0	0	0.0

Figure 1: *First 5 rows of the dataset.*

Representation of each class: Figure 2 shows the class distribution, which as we can see is imbalanced. This is often the case in real business problems, and this means that accuracy is a poor choice as a performance metric. The reason is that accuracy gives high scores to models which just predict the most frequent class. F1 score would be a better choice in this case.

Potential problems:

- Spelling mistakes (e.g. Ampliaciin, mas.habitaciones)
- Inadequate labels (e.g. "Compra de inmobiliario de oficina" labeled as "renta")
- Some proceeds dealt on more than one category but had only one label.

These mistakes could harm the ability to make predictions, because the models would learn the wrong labels. However, before dedicating time and effort to correct these mistakes, it is important to estimate whether this would make a significant improvement in performance. This requires a thorough error analysis (Describir cómo se hace)

Deep Learning algorithms are very robust to random errors in the training set, provided the number of errors is small. However, if these errors are systematic, the model could learn to misclassify.

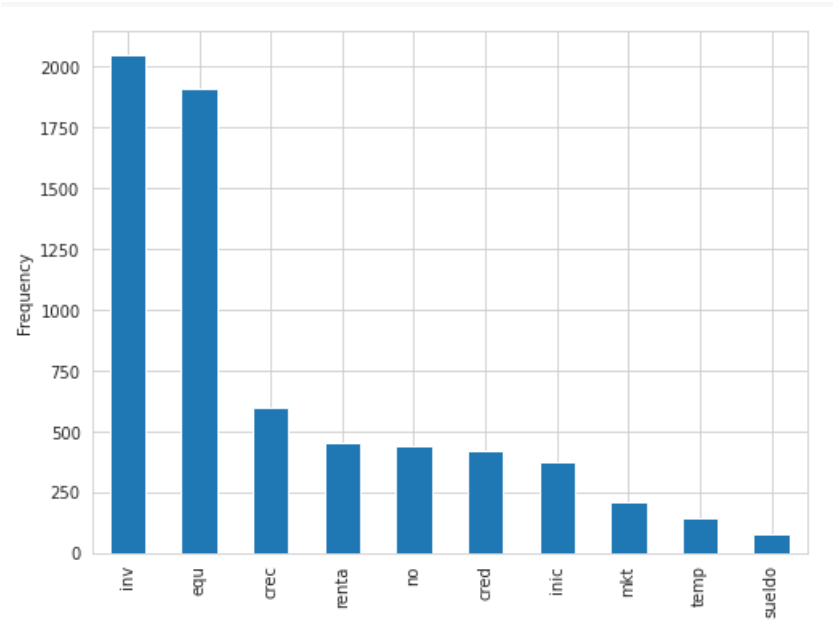

Figure 2: *Class distribution.*

3 Data Preparation

Data cleaning: Ver cómo cambia esto en ML

The variable "temp" had 6 missing values, which were imputed using the median (0). Another option could have been to drop those rows, but since most of them did have other labels, it was considered best to keep them. Also this variable vas of type float when it should be int, so it was converted accordingly.

Text processing: The text was pre-processed as follows:

- All text was converted to lower case.
- The following symbols were replaced by space:

  `/(){}[]|@,;`

- All symbols that are not digits, letters, space, or the following, were removed:

  `#+-`

- Spanish stop words were removed.
- Digits in text were removed (e.g. mercade5ria replaced by mercaderia)

Tokenization: Text was tokenized using Keras's Tokenizer class.

4 Modeling

We used the LSTM model (Long Short-Term Memory), which is a Recurrent Neural Network architecture.

Alguna figura de Andrew?

The first layer is the embedding layer that uses 100 length vectors to represent each word. The second layer is SpatialDropout1D, which performs variational dropout. The third layer is the LSTM layer with 100 memory units. The output layer (dense) creates 10 output values, one for each class. Activation function is softmax for multi-class classification. Because it is a multi-class classification problem, categorical crossentropy is used as the loss function.

Por qué elegí este modelo

5 Evaluation

Usar F1, y actualizar las figuras

Figures 3 and 4 show the accuracy and loss curves respectively. On figure 3 we can see strong overfitting (A qué se debe, cómo solucionarlo). Figure 4 shows that loss is going up on the test set. This is another indicator that the model is overfitting. The following steps are reccomended:

- Reduce model capacity.
- Add regularization.
- Check that the training and test splits are statistically equivalent.

Probar estas cosas!

6 Conclusion

Agregar la parte de ML!!

Resumen del performance de los distintos modelos, causas y cómo solucionarlo.

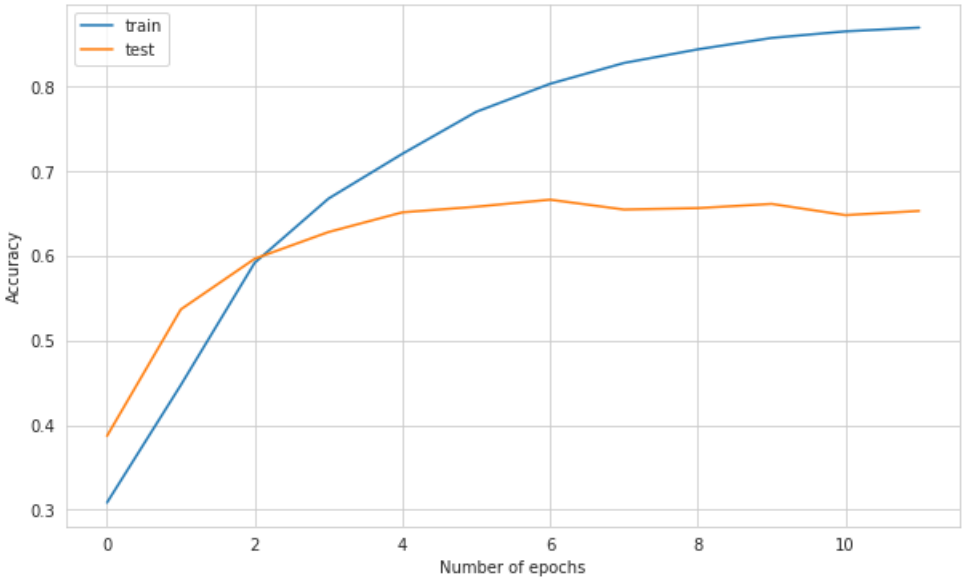

Figure 3: *Accuracy curve.*

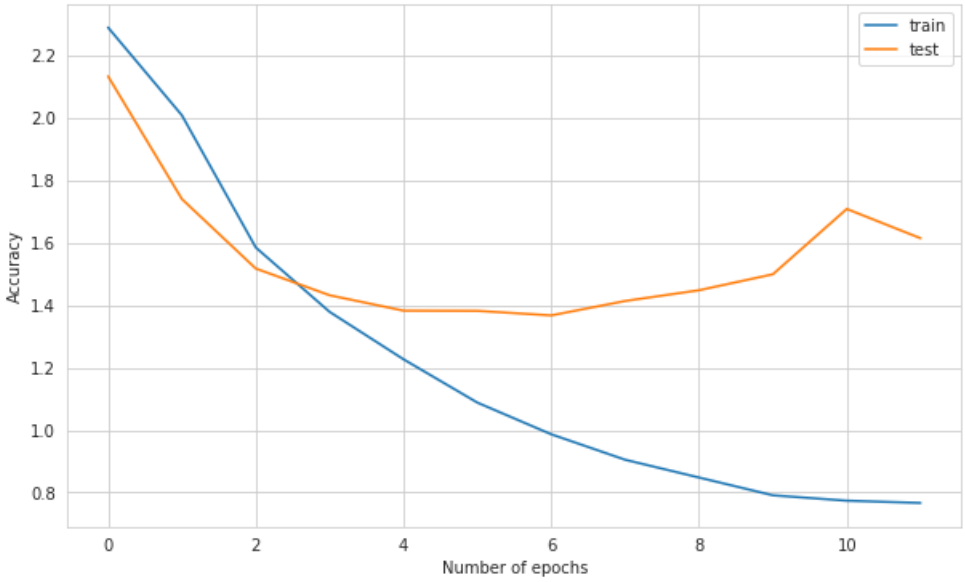

Figure 4: *Loss curve.*