Document Classification

AI approach

Stan Taov  
 Ontario, Canada  
 stastaov@gmail.com

Luiz Augusto de Carvalho  
 Ontario Canada  
 ziulcarvalho@gmail.com

FirstName Surname  
 Department Name  
 Institution/University Name  
 City State Country  
 email@email.com

ABSTRACT

Large part of the clerical work in this accountant organization consists of checking documents, classifying them and extracting some basic data from them.

This process today is basically manual, and therefore time consuming and prone to human error.

The idea behind this project was to create a code capable of based on the scanned PDF files classify the documents into six types:

* Invoices (Notas in Portuguese)
* Personnel payments (RPA in Portuguese)
* Receipts (Recibos in Portuguese)
* Bank transfers (DOCs in Portuguese)
* Bills (Boletos in Portuguese)
* Others (Outros)

Note that the class others will encompass basically everything else not covered in the previous five.

Other important point is the fact that data will be extracted only from the three first classes. Here there is another difference the type of data extracted from each type has similarities but also differences:

Data and value will be extracted from all three types

Company number origin and destiny of the Invoice (Just the invoices)

Company paying number and social Insurance number of the person receiving (Just for Presonnel payments)

CCS

Computing methodologies, Artificial intelligence,

KEYWORDS

Mask Linear Regression, decision three, Random Forest,.

INTRODUCTION

We will work with a dataset of 15.731 pdf files containing 16.475 documents (We may have a pdf file with more than one page on it).

We had labels of 4.448 of these documents. The label document was identified as revisado6.csv and encompasses the name of the pdf file and its class.

ANALYTICAL SCHEMA

This work has as objective to create an AI model that based on the PDF documents converted to TXT classify them into six categories.

The first issue identified in the database provided was the unbalance between the Invoices and the other five classes:



This unbalance adds difficult to model the problem, we not only have few samples of some classes to work with.

To circumvent this limitation we are going to follow two strategies:

We worked the changes in the baseline model following three types of strategies:

Data augmentation (Oversample): Using data augmentation techniques we evaluate the impact of the increase of the number of samples in our training base in the quality of our predictions.

DATA EXPLORATION

The data was provided by Krypton BPO and encompasses data regards the whole years of 2019 and 2020. The Pdfs were forwarded as they were forwarded by their clients. That means we have all sorts of types os scanned methods: From pictures taken to upside down scanned doc. Very few documents were 100% aligned,

The original CSV file layout is as follows:

RangeIndex: 4448 entries, 0 to 4447

Data columns (total 2 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 name 4448 non-null object

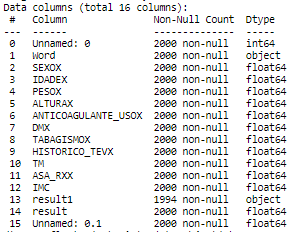
1 label 4443 non-null object

dtypes: object(2)

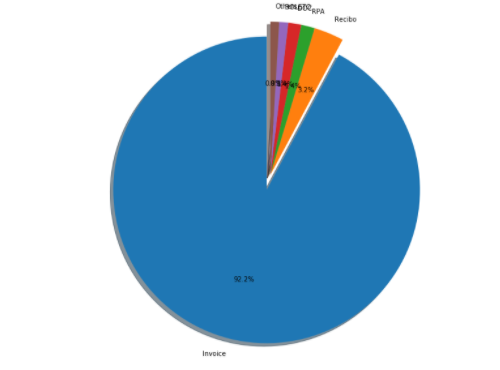
memory usage: 69.6+ KB

Five rows had Nan in the label

The treated database got the following layout:



Note that the columns 0,1,13 and 15 have no relevance in the analytical model being there only for documentation and indexing propose. The parameters used as reference for prediction are the ones between 2 and 12 and the output is identified in columns 14 (result).



DATA PREPARATION

We the labeling file we identified five rows where the labels were missing.

In this particular project the data preparation encompasses basically aligning the pdf files (almost all of them had some small inclination). We also found some very small number of pdfs which were scanned up side dow or sideways. The code doesn´t treat these cases.

To align the files we convert the pdf files into jpeg – using PIL library. Once aligned we convert the jpeg files inot TXT using py.tesseract OCR.

Once we have the file converted into text we clean it through 3 sequences of cleaning

1. Clean – Removes the special caracters
2. Clean1 – remove isolated words (Portuguese)

CREATE SYNTETIC VARIABLES

We create six synthetic variables which count how many times specific words appear in the document. Each class has its specific words:

DOC

['BANCO','FATURA','PROVANTE','VENCIMENTO','DEBIADA']

Invoces

['NFS','NOTA','FISCAL','PREFEITURA','MUNICIPAL']

Receipt

['RECIBO','RECEBEMOS','RECEBI','ALUGUEL']

Personnel payment

['RPA','NOMO','AUT','EMITENTE']

Bill

['BANCO','PROVANTE','BOT','DEBIADA','VENCIMENTO']

Others

['CONTRATO','AUXILIAR','DOCUMENTO','DANFPS']

Each time one of these words appear one of the six variables are added:

Sint1,Sint2,sint3,sint4,sint5,sint6

BASELINE MODEL

The first step is to develop a baseline model. In our specific case we are going to use as our basic model random forest

* Preparing the model
* Training the model
* Evaluating the model
* Making predicitions.

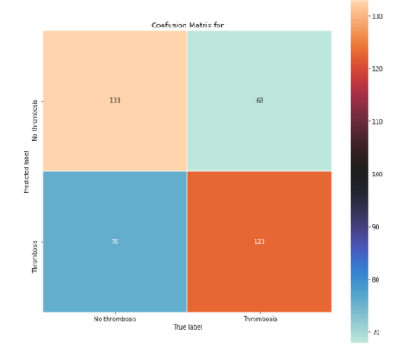
We are going to separate the oversampled database into training and test in a proportion of (80% training and 20% test). That means we are going to use xxx samples to train out model and apply this trained model to predict the output of 400 samples. Doing that we are going to check how many of these 400 the model predict correctly.

We are going to test several techniques:

* Linear Regression
* Decision tree
* Random Forest
* Ensembled combination of the previous 3

Once we identify which one of these models worked better we are going to adjust the hyper-parameters to improve the quality of the prediction even further.

Model 1 – Linear regression

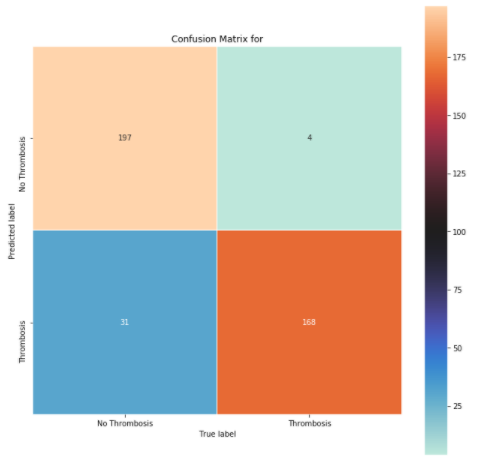


This model got it right 64% of the samples. Note that in this particular problem is important to get the false negative as smaller as possible. That means a good model may guesses the patient having thrombosis when in fact it hadn’t but what it cannot do is to predict that a patient is not going to have it and in the end he ends-up having.



Model 2 – Decision Tree

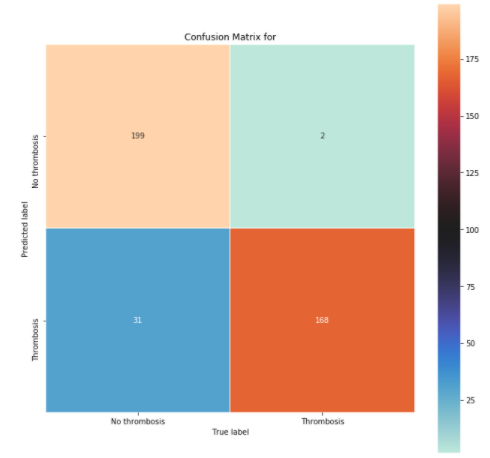
.



Note that this model seems to be much better than the previous one. It not only got a much higher percentage of right predictions but a very low percentage of false negatives.



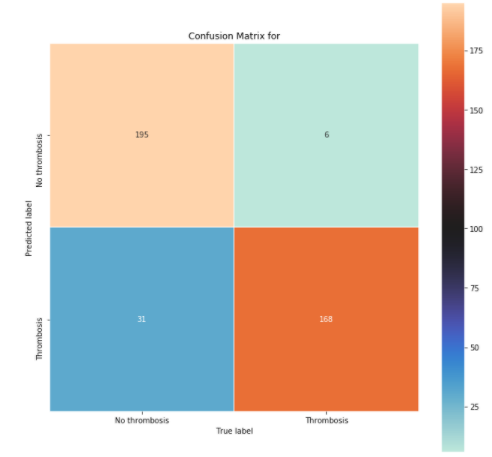
Model 3 – Random Forest



Note that we got a small but important improvement in the false-negative using this model (50% smaller). The number of false positive remained equal the previous model. Overall this model got right 0,50% more than the previous one but this improvement come from the fase-positives what in our context is very important.



Model 4 – SVC (Vector Classifier)



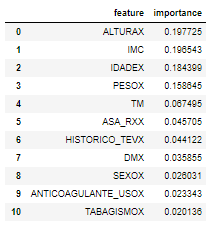
This model proved to be better than Linear regression but inferior than both Decision three and Random Forest.



Therefore becomes clear que the best model to predict the occurrence of thrombosis is Random Forest (Among the four models tested). It got 91.75% correct, 0.5% of false negatives 7.75& false positives. That means for each 200 patients the model will predict correctly 184 (having or not Thrombosis), predict incorrectly that 15 will have and predict incorrectly that 2 will not have.

THE IMPORTANCE OF THE PARAMETERS

Here we are going to evaluate the impact of each one of the parameters in the decision process of the model. This is an important evaluation because tells us what is really being important when taking the decision to classify a patient as “No thrombosis” or “Thrombosis”:

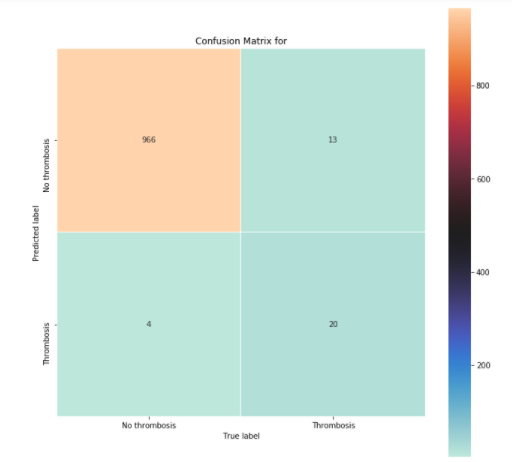


Here some very interesting things can be seen: IMC seems to be the most important factor in the definition (Considering that IMC is a composition of ALTURA and PESO (Height and Weight), followed by the age. A way down we have TM, SAA classification and History of TEV. Sex, use of anticoagulant and Tabagism seem to have almost no importance regards having or not Thrombosis.

Being old and over-weight seems to be the defining factors, having a pre-existing illness adds to your chance to having problems. Pretty common sense conclusion.

VALIDATION

We select a 50% subset of the original file (not balanced) and ask the trained model to classify it for us.



As we can see the results we a bit worse than the one obtained with the test set:

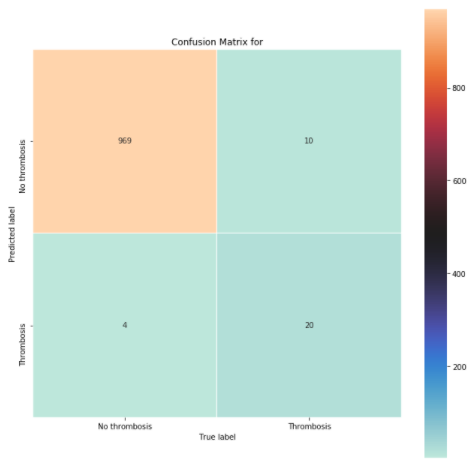


This is an indication that there is some overfitting in the training dataset, what is expected in a context where we had so few positive events to work with.

If we analyze the minority class (thrombosis) e can see that the model was capable of identifying 20 out of the 33 (60% of the events).

TRYING TO IMPROVE THE MODEL

One strategy to avoid overfitting is known “ensembling” it consists basically combining the models and by majority vote classify the item. Combining the four models we manage to archive a small improvement:





Small improvement in the total percentage of correct predictions but almost 30% reduction on the false negative.

If we think that the unbalanced class (thrombosis) is 2,24% of the samples that means the model gets it right 55% of the times.

REFERENCES

HCPUnet, Healthcare Cost and Utilization Project. Agency for Healthcare Research and Quality. Disponível em: http://hcupnet.ahrq.gov

Ethgen O, Bruière O, Richy F, Dardennes C, Reginster JY. Health related quality of life in total hip and total knee arthroplasty. A qualitative and systematic review of the literature. J Bone Joint Surg Am. v.86-A, n.5, p.963-974. 2004.

Khokhar A. et al. Venous thromboembolism and its prophylaxis in elective knee arthroplasty: An international perspective. The Knee. v. 20, p.170-176. 2013

Sivashankar Chandrasekaran, Siva Kumar Ariaretnam, Jason Tsung and David Dickison. Early mobilization after total knee replacement reduces the incidence of deep venous thrombosis. ANZ J Surg. v.79, p.526–529. 2009.

Anderson, M. Dunbar, J. Murnaghan, S.R. Kahn, P. Gross et al Aspirin or Rivaroxaban for VTE Prophylaxis after Hip or knee arthroplasty. N Engl J Med p.378:388. 2018

Andersen P,K, Kehlet H. Should deep venous thrombosis prophylaxis be used in fasttrack hip and knee replacement? Acta Orthopaedica. v.83, n. 2 p.105–106. 2012.

O’Reilly RF, Burgess I, Zicat BA. The prevalence of venous thromboembolism after hip and knee replacement surgery. Med J Aust. v.182, p.154-159. 2015.

Song K, Xu Z, Rong Z, Yang X, Yao Y, Shen Y, et al. The incidence of venous thromboembolism following total knee arthroplasty: a prospective study by using computed tomographic pulmonary angiography in combination with bilateral lower limb venography. Blood Coagul Fibrinolysis. v.27, n.3, p.266–269. 2016.

Quinlan DJ, Eikelboom J, Dahl O, et al. Association between asymptomatic deep vein thrombosis detected by venography and symptomatic venous thromboembolism in patients undergoing elective hip or knee surgery. J Thromb Haemost. v.5, p. 1438-1443. 2007.

Howie C, Hughes H, Watts AC. Venous thromboembolism associated with hip and knee replacement over a ten-year period: a population-based study. J Bone Joint Surg Br. v87-B p.1675–1680. 2005.

Zhang et al. Risk factors for venous thromboembolism of total hip arthroplasty and total knee arthroplasty: a systematic review of evidences in ten years. BMC Musculoskeletal Disorders. p.16-24. 2015.