HarvardX - PH125.9X - Data Science: Capstone - Information Access Prediction System in the Federal Executive Branch of Brazil

Flávia Lemos Xavier

15/08/2019

Abstract. This document is the project closure report for the course HarvardX - PH125.9x Data Science: Capstone. In this document, I describes my solution to develop a algorithm to predict the probability of access to information in the Federal Executive Branch of Brazil, based on jutrisprudential research, using the public dataset available in the Government's open data 1 (https://esic.cgu.gov.br/sistema/Relatorios/Anual/DownloadDados.aspx). I developed the algorithm using Logistic regression and Generalized Linear Models, once we have categorical data. This document consists by four sections: (1) **Introduction** describe the project background, goals, method and project steps, (2) **Analysis** presents data exploratory analysis, insights gained and the modeling approach, (3) **Result** presents the modeling results and discusses the model performance, (4) **Conclusion** describes a brief summary of the report, its limitations, and future work.

1. Introduction

The legal system is not as straightforward as coding. Just consider the complicated state of justice today, whether it be problems stemming from backlogged courts and overburdened public defenders.

Although the impacts of new technologies on society raise a number of ethical and legal issues, it is undeniable that artificial intelligence (AI) can facilitate the public decision-making for granting citizens rights.

So, can artificial intelligence help?

Very much so. Law firms are already using AI to more efficiently perform due diligence, conduct research and bill hours. But some expect the impact of AI to be much more transformational. 2 (https://www.forbes.com/sites/cognitiveworld/2019/02/09/will-a-i-put-lawyers-out-of-business/#6c27f32c31f0).

One of the applications of law that will certainly be supported by AI is jurisprudential research and probability analyzes of the public decisions, providing greater agility and precision. Of course, the accuracy of computational models must be based on high-value predictor, in other words, more context means better results.

This already occurs with the robot lawyer Ross. This robot was created by IBM and it is used by one of the largest US Law Firm with a Global Reach, Baker & Hostetler, to find and analyse the most relevant cases and laws. Thus, lawyers do not need to spend more time than necessary finding the applicable legislation and case law on the subject. 3 (https://rossintelligence.com/)

Similarly, artificial intelligence will accelerate the judicial and admionistrative process.

1.1 Background

As part of the team of the National Ombudsman´s Office (OGU), in the Federal Comptroller General's Office (CGU), I am responsible for the technical supervision and guidance of all ombudsman´s units in the Executive Branch on the federal level. Our team examine claims related to the delivery of public services; suggest

disciplinary measures and work to prevent faults and omissions of managers responsible for the inadequate delivery of public services. Additionally, we contribute to the dissemination of new forms of social participation in monitoring and supervising the delivery of public services; and promote capacity-building actions related to ombudsman´s activities. We also coordinate the Information Access System established by Law Nº 12,527/2012, in order to promote its good compliance throughout the Federal Executive Branch in Brazil.

1.2 Project objective

In this context, this project aims to initiate an algorithm development project to facilitate the implementation of the Law on Access to Information in the Federal Executive Branch.

In this way, I will create a pilot project to develop an algorithm that is capable of raising administrative jurisprudence and predicting as accurately as possible the decision-making tendency of all Federal Executive Branch bodies in relation to a given request category.

1.3 Method

The algorithm will initially be based on a categorical variable or predictor: the main subjects for access to information to a federal government agency, already identified and included in the initial database by subcategory of selected keywords in an Electronic Government Controlled Vocabulary (RequestCategory).

Thus, it will be necessary to use some principles and methods of Data Mining and Machine Learning such as the Logistic regression and the Generalized Linear Models 4 (https://rafalab.github.io/dsbook/). In both models, we show how the regression approach can be extended to categorical data. For binary data, we can simply assign numeric values of 0 and 1 to the outcomes y.

In this case the outcome is the probability that, in a given theme or situation, the government agency decide administratively favorable (y = yes/“Success” or “Acesso Concedido”, in Portuguese) or unfavorable (y = no/“Failure” or “Acesso Negado”) to the citizen request. If we define the outcome Y as 1 for “Acesso Concedido” and 0 for “Acesso Negado”, and
X as the RequestCategory, we are interested in the conditional probability:

$$Pr(Y = 1 \mid X = x)$$

Basically, I will train the algorithm with the models taking into account the existing databaset.

It is important to clarify that this project is an initial attempt to apply the course content. There is great potential for algorithm development by other predictors that include deeper contextual data, in addition to keywords such as word semantics, logical operators, requestor profile, etc.

In addition, I will take the opportunity of this project to practice practice skills on data wrangling, data visualization, machine learning, reporting using several packages, such as tidyverse, caret, ggplot2 and rmarkdown.

** 2. Analysis**

**2.1 Data Analysis

Data exploratory analysis: we will explore and visualize the data to have an overview with-in and between the variables, what’s insights gained after analysis. Main package for this step is tidyverse, to handle the cleaning, exploring and visualizing tasks.

The data file have `{r} nrow(data)` rows and 3 columns. There are 3 columns where:

- “IdSolicitante” - doubles: requester’s unique identifier;

- “Categoria do Pedido” - character : request category assigned by the Citizen Information System in accordance with the Electronic Government Controlled Vocabulary;
- “Tipo de Resposta” - character: administrative decision.

Top rows of data file

Id do Solicitante	Categoria do Pedido	Tipo de resposta
198	Governo eletronico	Acesso Concedido
229	Financas	Acesso Concedido
237	Recursos energeticos	Acesso Concedido
237	Defesa Nacional	Acesso Concedido
237	Defesa Nacional	Acesso Concedido
237	Defesa Nacional	Acesso Concedido

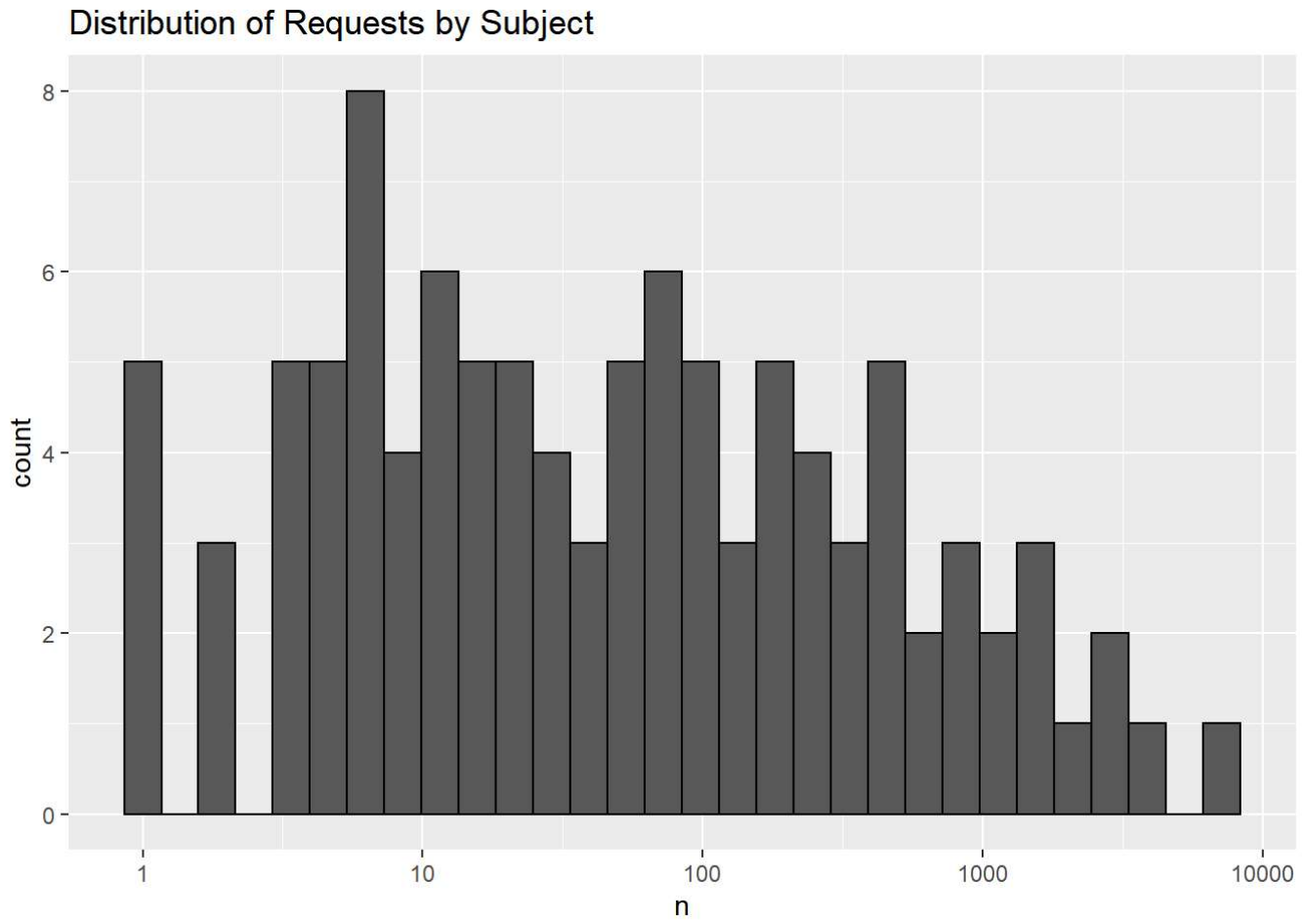
There are total 36,417 requests, from 17,220 different requesters, about 104 different subjects. In this database, once simplified, you can receive 3 responses from the federal government: “Favorable Decision” (“Acesso Concedido”, in Portuguese), “Partially Favorable Decision”(“Acesso Parcialmente Concedido”, in Portuguese) or “Denial Decision” (“Acesso Negado”, in Portuguese).

Summary of data (part 1)

number_of_rows	number_of_column	number_of_different_requesters	number_of_different_subjetcs	number_of_different_decisions
36417	3	17220	104	3

Note that each row is a request for access to information made until August 15, 2019.

We can identify below that the interest for each request category or request subject is quite uneven.



The most requested subjects to the Brazilian federal government in this first semester of 2019 are related to public administration itself, higher education, energy resources, finances and participation and social control in health.

```
## Selecting by count
```

Ranking of the most requested subjects

Categoria do Pedido	count
Administracao publica	7113
Educacao superior	3738
Recursos energeticos	2902
Financas	2469
Participacao e controle social em saude	2028

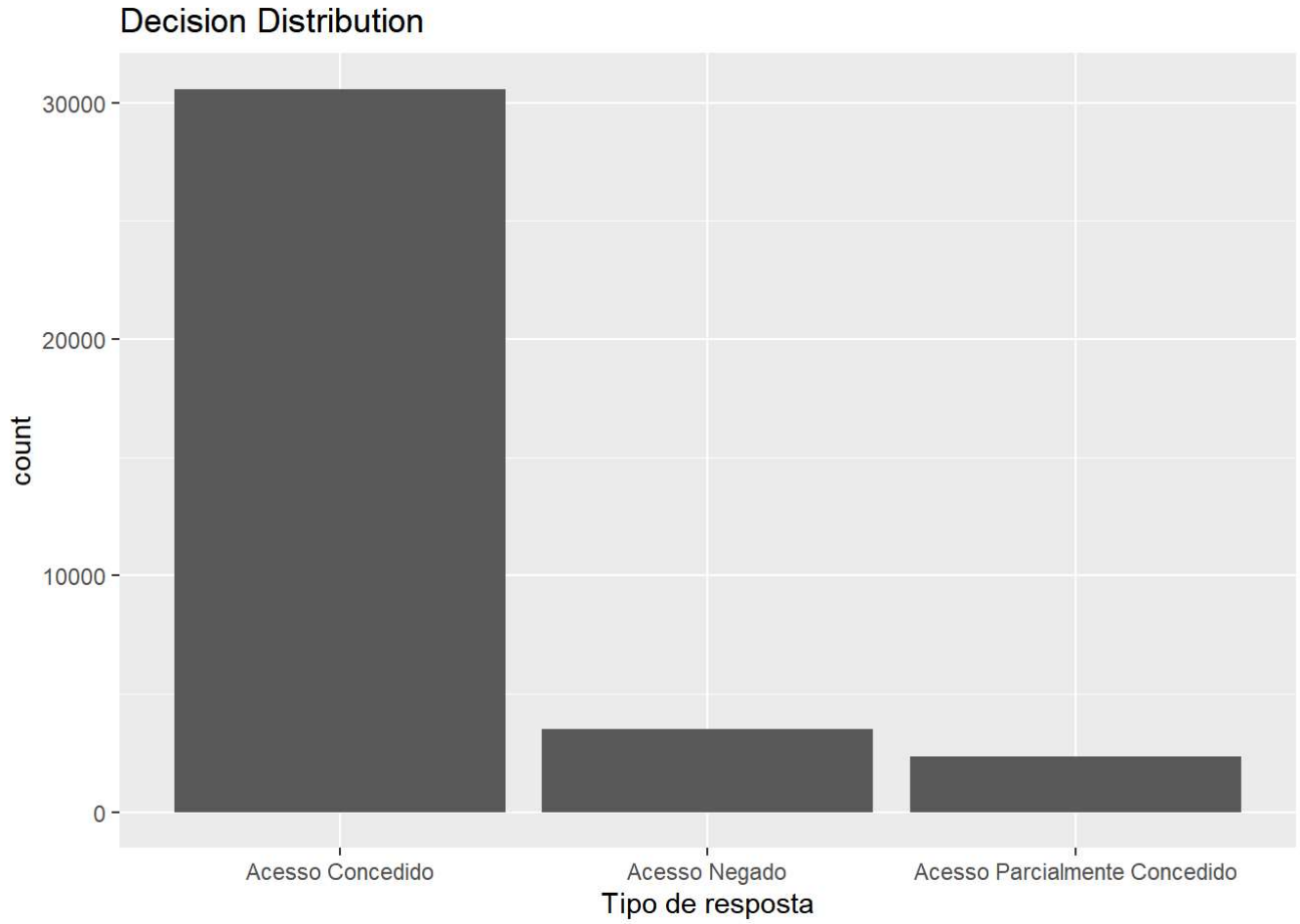
On the other hand, the least requested subjects to the federal government of Brazil in this first semester of 2019 refer to education for quilombolas, hospital infection, pipeline transportation, international transportation and violence.

Ranking of the least requested subjects

Categoria do Pedido	count
Educacao para quilombolas	1

Categoria do Pedido	count
Infeccao hospitalar	1
Transporte dutoviario	1
Transporte internacional	1
Violencia	1

Most of the requests received a favorable decision, followed well behind by denied decisions and even fewer by partially favorable decisions.



Now let's better understand whether each subject is likely to receive a favorable decision or not from the algorithm.

Firstly, we remove the id requester column because we won't use it as a predictor. Then, in a simplification, we also consider partially favorable decisions as favorable decisions.

```
data<- select(data,"RequestCategory"='Categoria do Pedido','Decision'='Tipo de resposta')

data$Decision[data$Decision == "Acesso Parcialmente Concedido"] <- "Acesso Concedido"

data<-data %>%
  mutate(Decision = factor(Decision, levels = c("Acesso Concedido", "Acesso Negado" )))

levels(data$Decision)
```

```
## [1] "Acesso Concedido" "Acesso Negado"
```

**2.2 Data Set

Now we can create the train and test data set to develop our algorithm.

The data set is generated by following code:

```
set.seed(1, sample.kind = "Rounding")
test_index <- createDataPartition(y = data$Decision, times = 1, p = 0.5, list = FALSE)
train <- data %>% slice(-test_index)
temp <- data %>% slice(test_index)
```

Now we make sure RequestCategory in validation set are also in train set and we ddd rows removed from validation set back into train set

```
validation <- temp %>%
  semi_join(data, by = "RequestCategory")
validation <-validation [-1,]

removed <- anti_join(temp, validation)
```

```
## Joining, by = c("RequestCategory", "Decision")
```

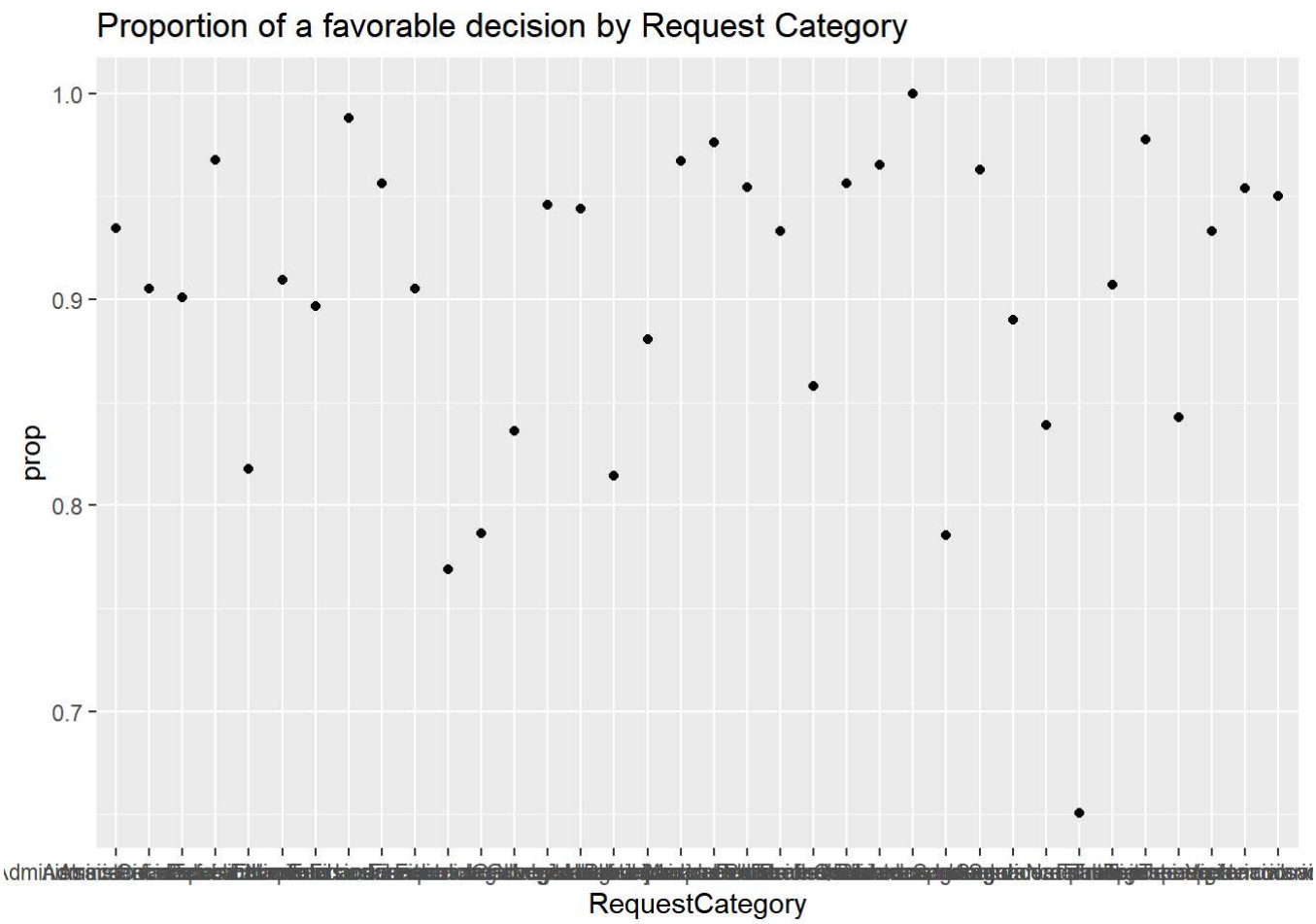
```
train<- rbind(train, removed)

y <-train$Decision
x <- train$RequestCategory
```

So for illustrative purposes, we try to predict the decision using the subject (category request) values as predictors.

Firstly, let's provide a prediction for a request on Public Administration. What is the conditional probability of being favorable decision if the request on Public Administration?

```
## # A tibble: 1 x 1
##   y_hat
##   <dbl>
## 1 0.910
```

To construct a prediction algorithm, we want to estimate the proportion of the favorable decision for any given request category X=x, which we write as the conditional probability described above. Here is a preview of this probability in the existing database:



Since the results from the plot above look close to linear, and it is a simple approach, we will try logistic regression.

Note: because $p0(x) = 1 - p1(x)$, we will only estimate $p1(x)$

```
lm_fit <- mutate(train, y = as.numeric(y == "Acesso Concedido")) %>% lm(y ~ x, data = .)
head(lm_fit$coefficients)
```

```
##                   (Intercept)         xAdministracao publica
##                    0.91864407                    -0.00860567
##                         xaguas        xAlimentacao e nutricao
##                    0.08135593                     0.08135593
##              xAmbiente e saude xAmbientes ocupados pelo homem
##                    0.08135593                     0.08135593
```

3. Result

Once we have estimates our coefficients, we can obtain an actual prediction. To form a prediction, we define a decision rule: predict “Acesso Concedido” (Success, in english) if $p(x) > 0.5$. We can compare our predictions to the outcomes using:

```
p_hat <- predict(lm_fit, validation)
y_hat <- ifelse(p_hat > 0.5, "Acesso Concedido", "Acesso Negado")%>%as.factor()
confusionMatrix(y_hat, validation$Decision)$overall["Accuracy"]
```

```
##  Accuracy
## 0.9042728
```

We see this method does substantially better than guessing since our accuracy is so high (> 0.9).

4. Conclusion

In this report we described a way to build up a jurisprudential probability algorithm to predict administrative decisions using the Citizen Information System database of the Federal Executive Branch in Brazil (we call as “data”). One predictor were used in our algorithm: the subcategory of selected keywords in an Electronic Government Controlled Vocabulary (we call as “RequestCategory”).

The final accuracy of our algorithm is > 0.9.

A better result could be achieved applying “Naive Bayes” or “Quadratic Discriminant Analysis (QDA)”, once they are more complex models of conditional probability, with few predictors. For instance, “Quadratic Discriminant Analysis (QDA)”it is a version of Naive Bayes in which we assume that the distributions of the conditional probability are multivariate normal.

However this is an opportunity to improve our algorithm in the future by including more predictors, that could also improve the results further.

Reference

[1] *“The Government’s open data: the Citizen Information System”*. Available on 2019-08-15. link (https://esic.cgu.gov.br/sistema/Relatorios/Anual/DownloadDados.aspx)

[2]*"Will A.I. Put Lawyers Out Of Business?*Available on 2019-08-15. link (https://www.forbes.com/sites/cognitiveworld/2019/02/09/will-a-i-put-lawyers-out-of-business/#6c27f32c31f0)

[3]*"Ross Intelligence* link (https://rossintelligence.com/)

[4]*“Introduction to Data Science - Data Analysis and Prediction Algorithms with R”*, Dr. Rafael A. Irizarry link (https://rafalab.github.io/dsbook/)

[5] *R Markdown: The Definitive Guide*, Yihui Xie, J. J. Allaire, Garrett Grolemund, 2019-06-03 link (https://bookdown.org/yihui/rmarkdown/)