FELIPE
CEZAR

Figthing Corruption wih Machine Learning

Detecting Campaign Financing infractions In The 2018 Brazilian Elections

Corruption in the Age of AI

Corruption is a serious global problem, transcending borders, political affiliations and systems of government. As more governments digitize their operations, there is an increased opportunity to develop practical anti-corruption tools based on big data, open data and artificial intelligence

US$1 trillion

Is the amount of government revenues that the IMF estimates to be lost every year globally with corruption

Why Brazil?

The combination of high availability of data and high levels of perceived corruption make Brazil a particularly attractive case for the application of AI for anti-corruption efforts.

High Open Data Ranking

Brazil is ranked 8th best in the World

Data from Global Open Data Index
https://index.okfn.org/

High Perception of Corruption

Rated 4.6 out 5 in Perceived Corruption in Public Sector

Data from Our World In Data
https://ourworldindata.org/corruption

Project Summary

Machine Learning Model trained with data downloaded and scrapped from Tribunal Superior Eleitoral (Superior Electoral Tribunal) websites

Around 26k candidacies for President, Governor, Senator, Member of Congress were used to train the model.

The goal of the model is to predict whether a given politician should be investigated for financial infractions

Data Overview

Candidates Assets

Campaign Finance

Personal and
Candidacy information

Election Results

Court Cases

Personal and Candidacy Information

The data that identifies each candidate. It includes information such as:

- Home state
- Age
- Gender
- Party affiliation and partnerships
- Ethnicity
- Professional, and education background

Candidates' Assets

The data lists a candidates' total asset value as reported to the tax authorities

It also breaks down the total value by categories such as:

- Real Estate
- Onshore and offshore accounts
- Shares
- Jewelry, art collection

Campaign Finance

The data lists every inflow and outflow of funds reported for the campaign

In addition to the cashflows value, it also categorizes them in a variety of ways, including:

- Origin of funds
- Profile of donors
- Type of receipt provided
- Type of product or service provided

Election Results

The data lists the results from the election, pointing out who was elected for what, and how many votes they received

In addition to 2018, results from the 2014 and 2010 elections were also included to differentiate candidates based on previous performance

Court Cases

The data lists all court cases related to the 2018 elections, not just the ones about campaign finances

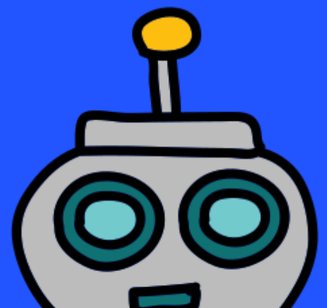
It shows which people and legal entities are involved in the case as well as the subject matter of the case

Machine Learning

1. Once all the data is obtained and cleaned, a model of supervised learning is built.

2. We indicate to the model which campaigns were involved in an enquiry about financial infractions, it then learns the patterns associated with them

3. Once it has been trained on a share of the data, it moves on to test its assumption on a different share. This is when we see how well it performs



Model Overview

- A **XGBoost** model was trained to classify each campaign as "flagged" or "not flagged" for possible infractions
    - **Out of** a total of **26k campaigns** in the dataset, **8.5k** were involved in a **court case related to their finances**.
- It was **trained** on around 18.2k datapoints and **tested** on 7.8k (**70%/30%**)

92.1%

accuracy

for all combined predictions

84.3%

of actual infractions

were flagged

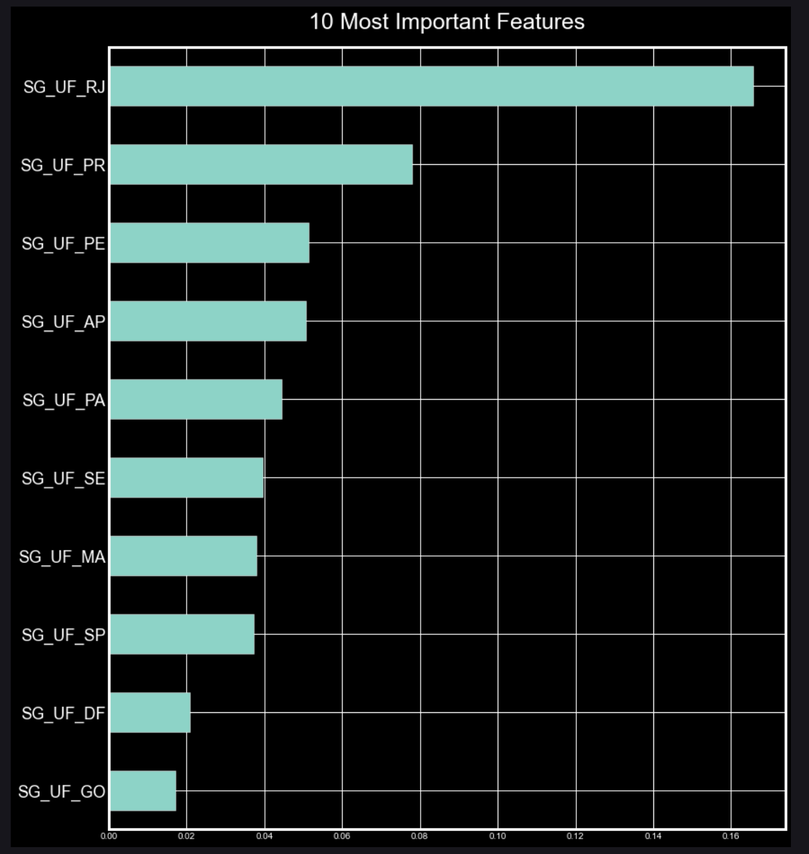
91%

of flagged campaigns

had been involved in a court case

Most Important Features

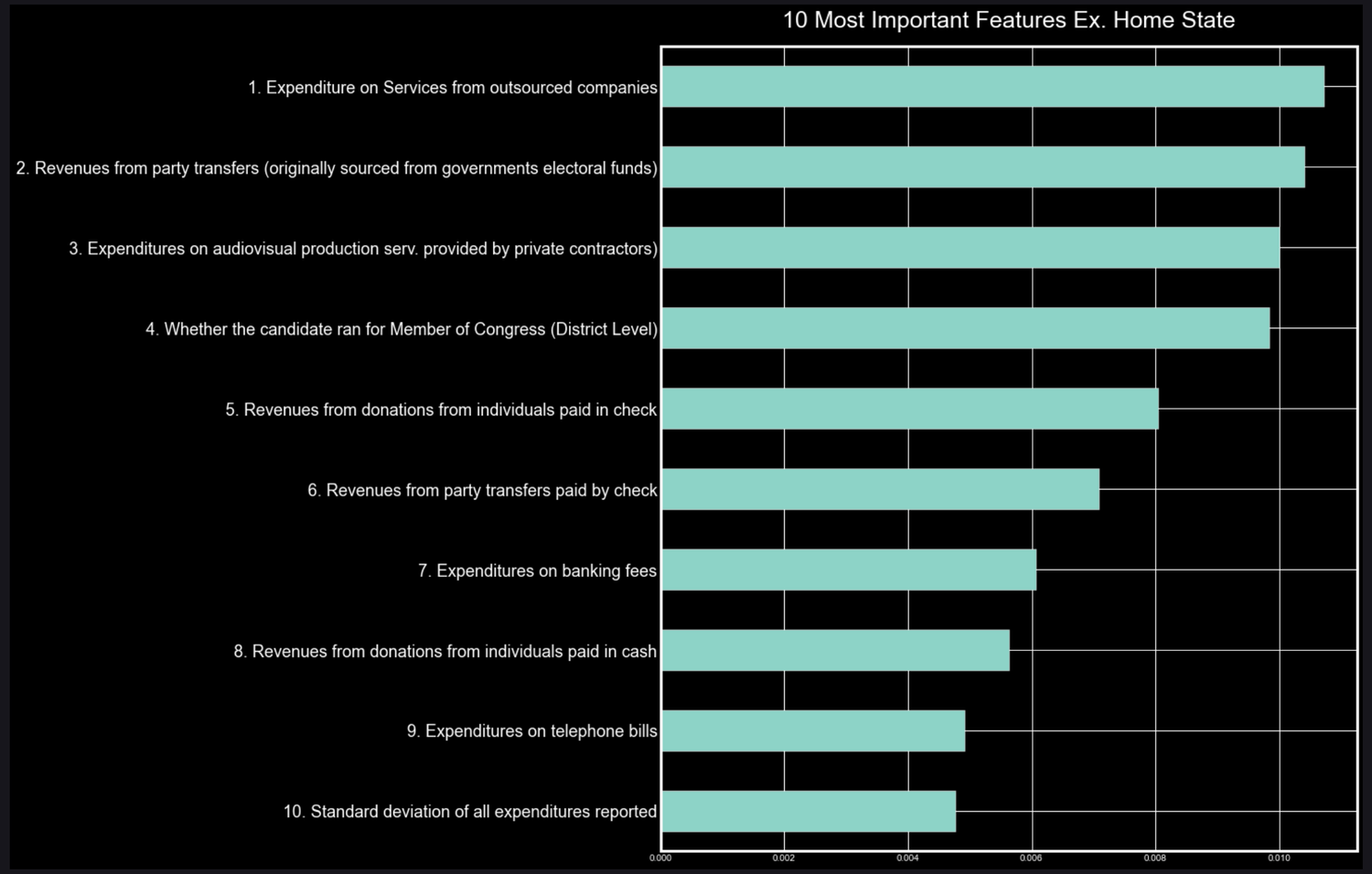
The importance score of a feature relates to how much that attribute is used to make key decisions with decision trees.



- The **State** where the campaign was held was found to be the **most important** features in determining whether the campaign should be flagged for possible infraction

Most Important Features

exc. Home States



- Once we exclude the States from the list, we find that it the amounts spent or received in specific ways are the most important features.
- For example, ranked as number 1 in the chart below we see the amount spent in services from outsourced companies. Following that we see amounts of revenues received as donations from other political parties, sourced from government electoral funds.

Further Research

- Explore court cases with different subject matter
- Reclassify court cases with a Natural Language Processing and clustering models in order to use them as new infraction flags
- Build unsupervised learning models that look for anomalies in the data

Thank you!

fc.felipecezar@gmail.com

https://www.linkedin.com/in/felipecezar1/?locale=en_US