**ESASA:**

Company Profile / Quick Facts:  
- Mexican PYME (~200 employees)  
- 33 years in the B2B Food Services ( Comedores ) Market  
- Geographic reach: CDMX Metropolitan Area & Toluca\*

- Roughly 70/30% Females-Males worker gender

- 90/10% Operative vs Administrative Profile

- Fierce recruitment competition (against ALL service industries

|  |  |  |
| --- | --- | --- |
| **Fiscal Year** | **Employees** | **Total Departments** |
| **2018** | **~230** | **35** |
| **2019 YTD** | **~200** | **32** |

**Problem Worth Solving:**

"Single / standalone business variable that could provide breakthrough in HR and other business variables"

**Decompose the Ask:**

Rotation 5Ws2H: What, when, where, WHY?, WHO, how, HOW MUCH/MANY?

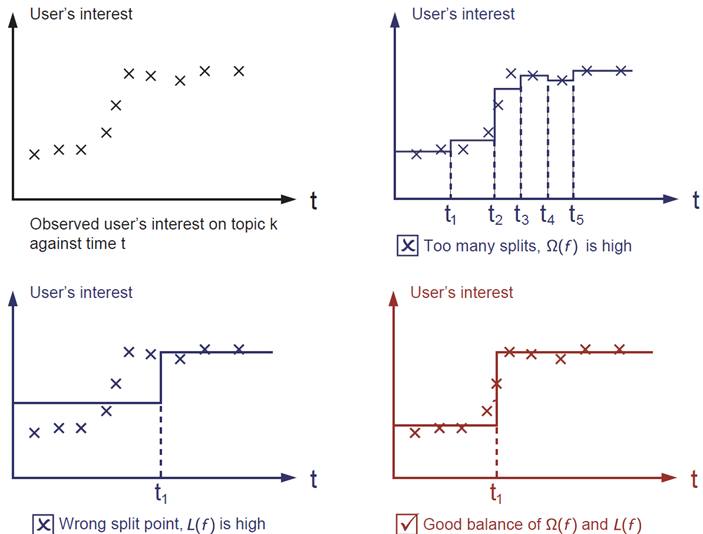
**Data/Sources (Strategy/Metrics) vs Analysis Needs and Scope:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Data Category** | **Scope** | **Source** | **Comments** |
| Detailed | ($, distribution, taxes/deductions, etc) | ESASA |  |
| Employee Profile | Basic: gender, age, address) | ESASA |  |
| Employee Job Satisfaction |  | ESASA | Availability vs Time constraints |
| Socioeconomic indexes |  | INEGI | Marginality Index used |
| Travel time 2 Work Center |  |  | Geopoints processing |
| Insecurity (work center) |  |  | Hoyodeldiablo API |
| Insecurity (trajectory) |  |  | Couldn’t be obtained |

**XGBoost,** <https://xgboost.readthedocs.io/en/latest/index.html>

### Function Objective: Training Loss + Regularization

The **regularization term** is what people usually forget to add. The regularization term controls the complexity of the model, which helps us to avoid overfitting. The general principle is we want both a simple and predictive model. The tradeoff between the two is also referred as **bias-variance tradeoff** in machine learning



**Main objective in the project:**

* To analyze the data with XGB in order to compare it’s output against the other models regarding prediction of employee status
  + X (vector of features) in this case, we have a dataframe with 450 rows and 20 columns (features).

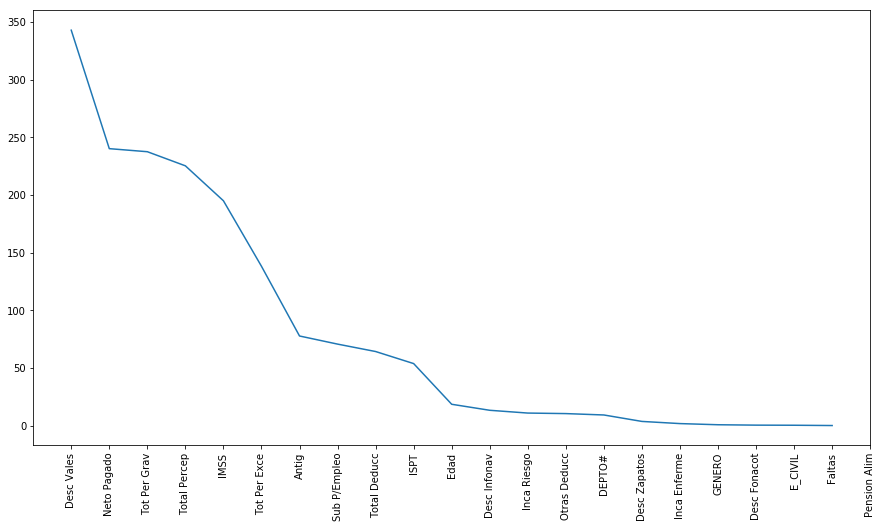
**Analysis Process**

**1. A first/ early draft round** was run with just 2019 cumulative data from ESASA w/o other external identified sources, to basically learn and understand the library while the final whole data was available. Train/test split data from sklearn were used for models 1 through 4

<https://github.com/mmastermind/HR_Boost/blob/XGBoost/HR_Boost/HR_Boost/Deep_learning_acumulados-Copy1.ipynb>

**2. A second round** using all the 2018 and 2019 ONLY payroll data was ran but selecting just 11 from the available 24 payroll columns using Univariate Feature Selection (UFS)from SKLearn. This library was used to understand the significance of all the data plugged given that the nature of the payroll components could be redundant for the model.

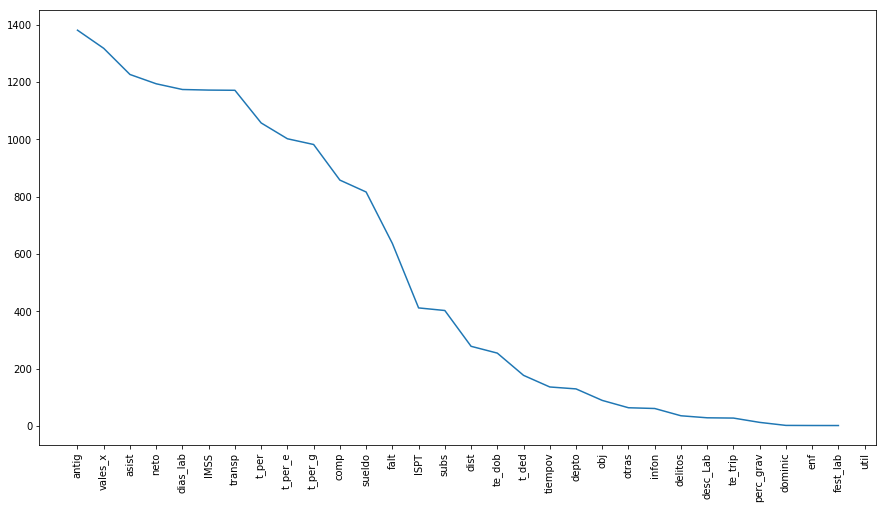
<https://github.com/mmastermind/HR_Boost/blob/XGBoost/HR_Boost/HR_Boost/Deep_learning_acumulados-Copy2.ipynb>



*Variable importance according to univariate feature selection (UFS), round 2*

Accuracy and features importance results are still not considered relevant since “external variables” (EV for short from now on, referring to travel time to work center and crime) are not included

**3. The third round** involved the full dataset and can be found in: <https://github.com/mmastermind/HR_Boost/blob/XGBoost/HR_Boost/HR_Boost/Deep_learning_acumulados-Copy3.ipynb> and showed the following:

**

*Variable importance according to univariate feature selection (UFS), round 3*

|  |  |
| --- | --- |
| **Feature** | **Importance** |
| antig | 0.205291 |
| neto | 0.116704 |
| vales\_x | 0.093870 |
| tiempov | 0.075123 |
| sueldo | 0.072301 |
| dist | 0.070555 |
| t\_per\_g | 0.067044 |
| t\_ded | 0.061931 |
| falt | 0.056249 |
| comp | 0.049539 |
| IMSS | 0.046739 |
| t\_per\_e | 0.029114 |
| ISPT | 0.026954 |
| te\_dob | 0.017243 |
| transp | 0.007339 |
| subs | 0.003561 |
| asist | 0.000443 |
| dias\_lab | 0.000000 |
| t\_per | 0.000000 |

After selecting the first 19 features according to visual analysis from this UFS and processing the model with XGB Classifier as in the past two rounds, the results were the following:

**Accuracy: 83.89% (Test Data)**

XGB also has the advantage to analyze the importance of features involved for the event, which in this 3.1 round were:

After studying this output in detail, a hypothesis came out about the “interference” of some of these features competing between one another in importance, when in the reality are in the same “category” (i.e. all income for the employee).

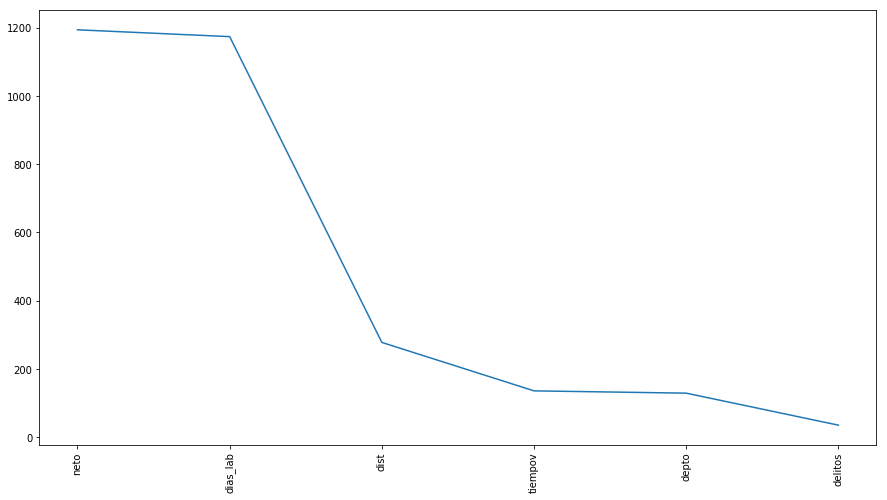
Thus, a sub-round within this model was run but using “interaction constraints”2 parameter from XGB which, in short, allows us to “bundle” some features in order to analyzing as one against others. In this case the bundled features were:

**params\_constrained['interaction\_constraints'] = '[1,2,3,6,7,8,9,10,11,14,16], [5,12,13,17]'**

With these constraints, the output of the same data and model was:

**Accuracy: 93.97% (Test Data,interaction constraints)**

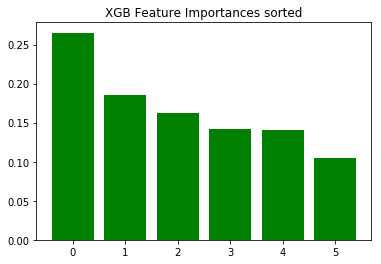
**4. Fourth round,** with the immediate previous’ results, I decided to run a model but discarding most of the previous’ features by simplifying the categories on: income, taxes (Deductions), crime, time to work center, days worked, distance to work center and department. Results were:



|  |  |
| --- | --- |
| **Feature** | **Importance** |
| **neto** | 0.265103 |
| **tiempov** | 0.185025 |
| **delitos** | 0.162770 |
| **dist** | 0.142362 |
| **depto** | 0.140559 |
| **dias\_lab** | 0.10418 |

**Accuracy: 80.18%**

<https://github.com/mmastermind/HR_Boost/blob/XGBoost/HR_Boost/HR_Boost/Deep_learning_acumulados-Copy4.ipynb>



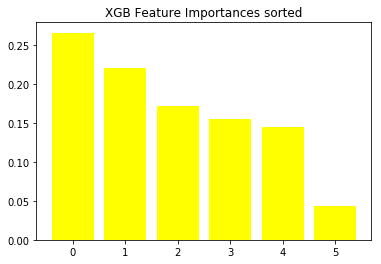
While accuracy decreased, from my standpoint using this output, in the worst case, it’s far more actionable in terms of actions, and more consistent with the survivability and neural\_networks models, because it provides more relevant information that later can be applied considering other factors (implied in previous-more detailed columns) such as payroll structure and taxes

**5. Final round,** based on round 4, this round was run under a different approach: to train the model with 2018 data and test it with 2019, under the assumption that more current data (2019) is more relevant than older (2018). Results were:

|  |  |
| --- | --- |
| Feature | **importance** |
| **neto** | 0.265738 |
| **tiempov** | 0.219876 |
| **delitos** | 0.171607 |
| **dist** | 0.155279 |
| **depto** | 0.144189 |
| **dias\_lab** | 0.043310 |

**Accuracy: 85.86% (Train)**

**Accuracy: 50.63% (Test)**



<https://github.com/mmastermind/HR_Boost/blob/XGBoost/HR_Boost/HR_Boost/Deep_learning_acumulados-Copy5.ipynb>

Further analysis on this line of thinking is presented within the approach of neural\_networks