ITBA
Instituto Tecnológico de Buenos Aires

Trabajo Práctico Final
82.05 Análisis Predictivo

– Abril Noguera

2022 1Q

AGENDA

01 Introducción

¿De qué se trata la base?

02 Objetivo

¿Qué se quiere predecir?

03 Análisis Exploratorio

Inspección y preparación de la base. Tratamiento estadístico y gráfico de los datos.

04 Hipótesis y Supuestos

Qué suposiciones existen sobre el análisis. Planteo del modelo.

05 Modelos de Predicción

Presentación de los modelos predictivos utilizados y explicación del modelo con mejor ajuste.

06 Fitting

Justificación del fitting del modelo.

01 Introducción

Encuesta de Inserción Laboral de Graduados Universitarios

- **Ámbito Poblacional:** graduados del sistema universitario español.
- **Ámbito Geográfico:** todo el territorio español, titulados de universidades españolas.
- **Ámbito Temporal:** se realizó la encuesta en el año 2019 con graduados del 2013 / 2014.

Instituto Nacional de Estadística

01 Introducción

Encuesta:

- Datos Personales y Sociodemográficos.
- Educación y Aprendizaje.
- Movilidad.
- Situación Laboral Actual del Graduado.

Instituto Nacional de Estadística

02 Objetivo

Objetivo General de la INe: *"El objetivo principal es conocer la situación laboral de los graduados universitarios, así como los diversos aspectos de su proceso de inserción laboral, es decir, el acceso al mercado de trabajo."*

Predecir el comportamiento laboral de los graduados universitarios.

- ¿Se puede predecir la inserción laboral?
- ¿Los graduados ocupan puestos acordes a sus estudios?

Preguntas Extra.

- ¿A mayor cantidad de estudios mayor sueldo?
- ¿La rama de estudios describe el sueldo?

03 Análisis Exploratorio

Variables Categoricas:

- ➔ Variables Nominales
- ➔ Variables Ordinales

Grupos:

- ➔ Personales del Graduado:
  - ◆ Sexo
  - ◆ Edad
  - ◆ Nacionalidad
  - ◆ Tipo de Hogar
- ➔ Estudios:
  - ◆ Rama de Estudio
  - ◆ Becado
  - ◆ Estudio en el Extranjero
  - ◆ Motivo
  - ◆ Capacidades
  - ◆ Otros Estudios
- ➔ Laborales:
  - ◆ **Situación Laboral Actual**
  - ◆ Situación Profesional Actual
  - ◆ **Nivel de Formación adecuado para el Trabajo.**
  - ◆ Área de Estudio apropiada para el Trabajo.
  - ◆ Sueldo
  - ◆ Intento conseguir otro trabajo?

Estas variables van a definir el Target, pero no serán utilizadas para el modelo.

03 Análisis Exploratorio

Registros Vacios:

- Variables que solo describen a una agrupación.
  - Ej: solo si sos empleado tenes sueldo…
- Vacíos sin sentido alguno.

Se agrega un valor a la variable que represente al otro grupo.

Se eliminan de la base por ser poco representativos.

Inconsistencias:

- No se encontraron valores inconsistentes, es decir que no cumplían con las respuestas predeterminadas.

Outliers:

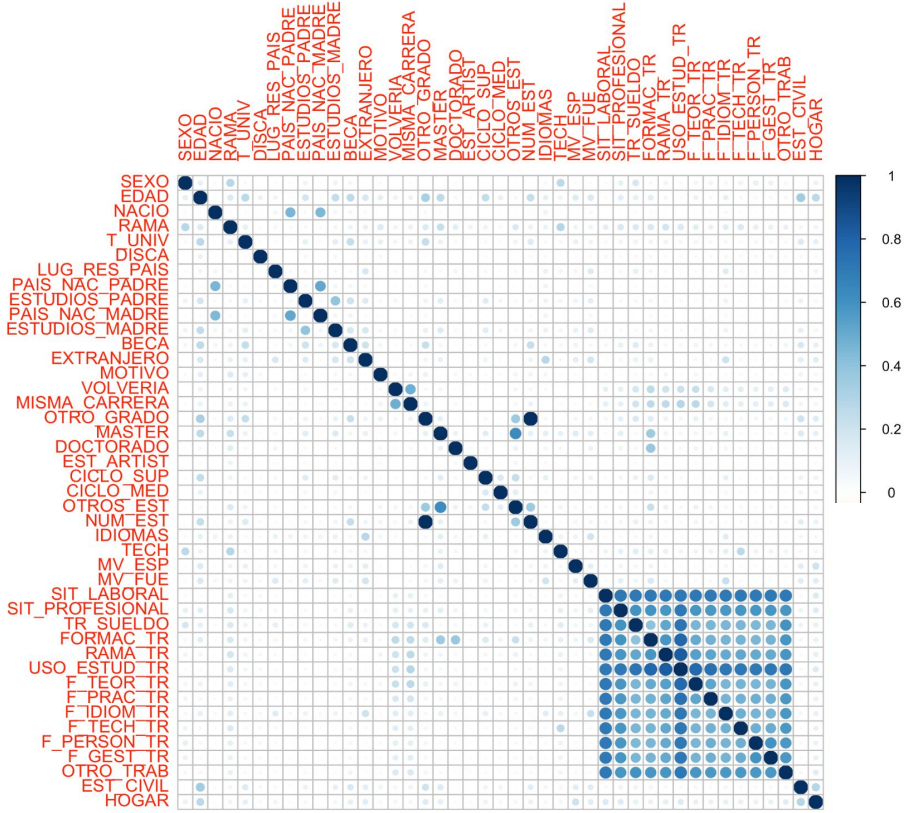
- No se encontraron outliers, porque se tratan de variables categóricas con respuestas predeterminadas.

03 Análisis Exploratorio

03 Análisis Exploratorio

Correlación de Cramer:

* Porque son variables categóricas. Mide la correlación de 0 a 1.

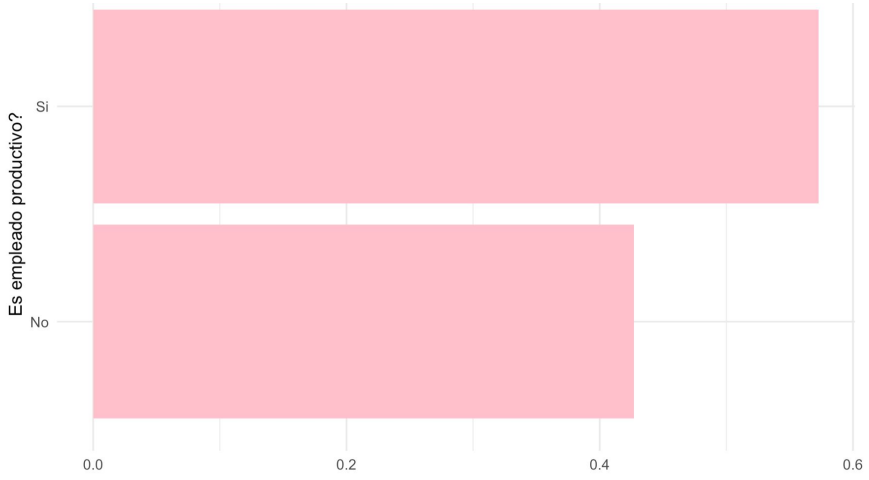


03 Análisis Exploratorio

Objetivo: Predecir el comportamiento laboral de los graduados universitarios.

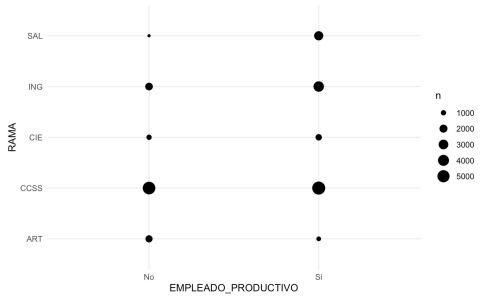
¿Nos interesa un graduado empleado pero qué no aplique su título?

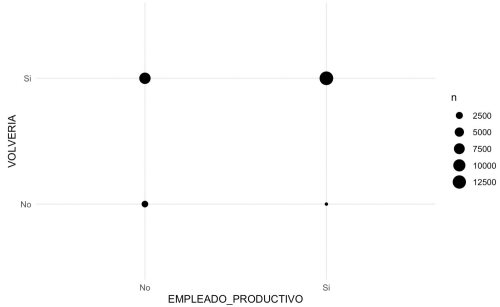
Empleado Productivo: graduado actualmente empleado que aplica sus estudios universitarios en su área en el trabajo al que se dedica.

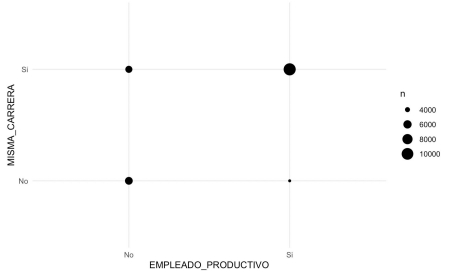


03 Análisis Exploratorio

Correlación de Cramer:







04 Hipótesis y Supuestos

¿A mayor cantidad de estudios mayor sueldo?

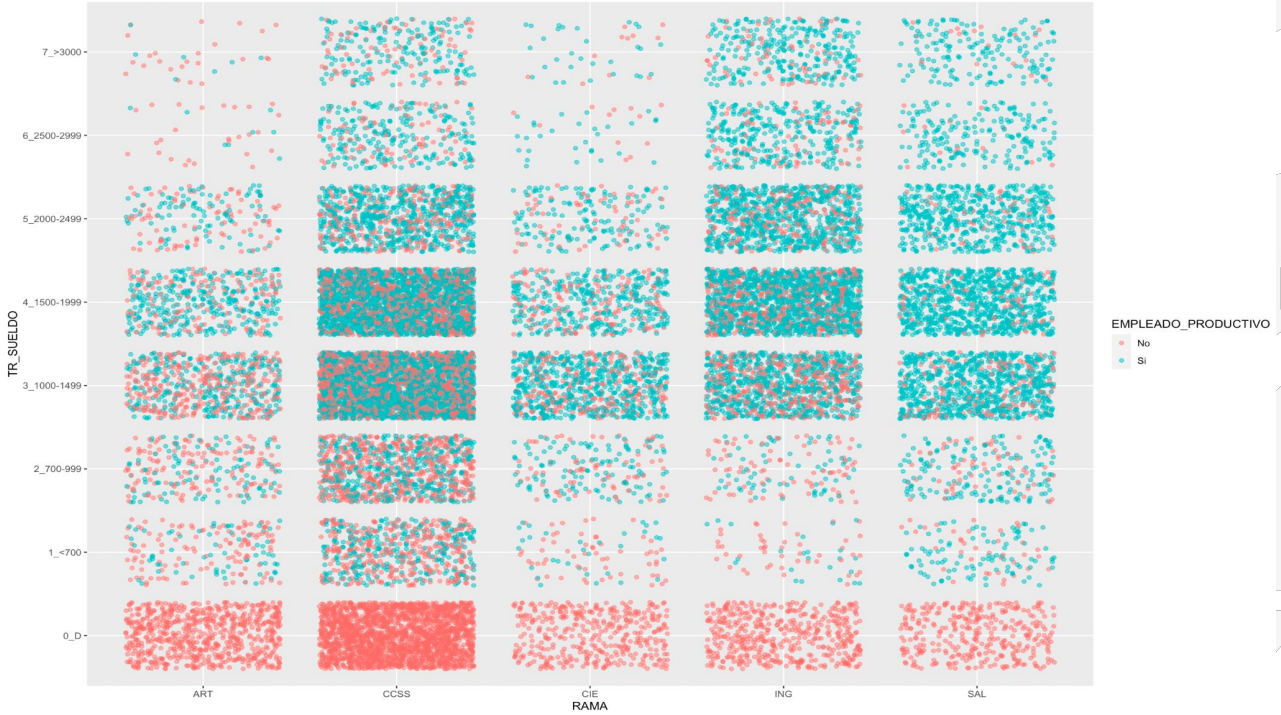
Correlación de Spearman:
13.64%
* Cómo son variables ordinales spearman tiene sentido. Mide la correlación entre -1 y 1.

Correlación de Cramer:
9.65%
* Porque son variables categóricas. Mide la correlación de 0 a 1.

04 Hipótesis y Supuestos

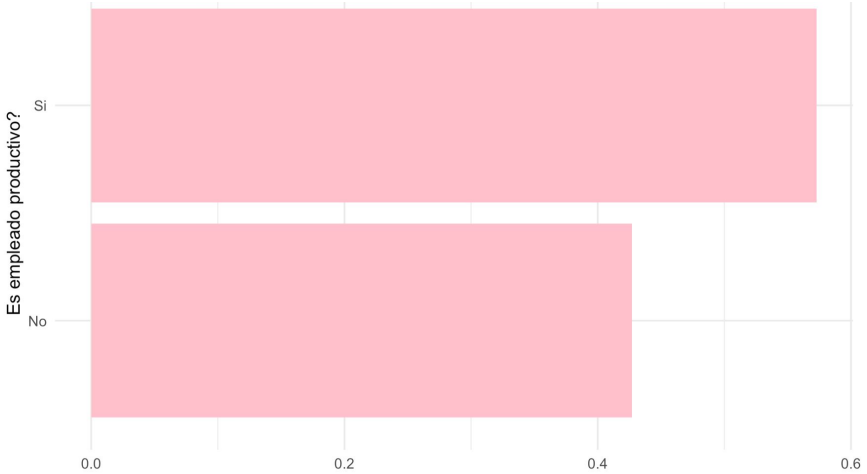
¿La rama de estudios describe el sueldo?

Correlación de Cramer:
15.22%



04 Hipótesis y Supuestos

- ¿Los graduados ocupan puestos acordes a sus estudios?



- ¿Se puede predecir la inserción laboral?
- ¿Existe un patrón que refleje el comportamiento laboral de los graduados?

BUSCAMOS QUE LO RESPONDA EL MODELO DE PREDICCIÓN

04 Hipótesis y Supuestos

Casos de Aplicación del Modelo:

¿Cuántos graduados aplicarán sus estudios?

¿Qué tipo de cursos o carreras le recomiendo a este graduado?
¿Qué está pasando con los planes de estudio para qué los graduados no los apliquen?

¿Qué tipo de empleo le recomiendo a este graduado?

05 Modelos de Predicción

Objetivo: Predecir el comportamiento laboral de los graduados universitarios.

Variable Target: Empleados Productivos.

Modelo: Clasificación.

Partición de la Base:

70% de Training
Stratified Split: por Rama de Estudio.

05 Modelos de Predicción

Métricas de Evaluación:

- Misma importancia a las categorías.
- Categorias balanceadas

Accuracy

AUROC

05 Modelos de Predicción

- Arboles:
  - **CatBoost**
  - Random Forest
  - AdaBoost
  - Decisión Tree
  - LightGBM
  - XGBoost
- KNN
- SVM
- Kmeans

Herramientas:

Grid Search

Cross Validation

One Hot Encoding

Feature Importance

PCA

Cat por categorías y Boost por qué usa Gradient Boosting.

- ➢ Bueno para variables categóricas.
- ➢ Bueno para información limitada.
- ➢ Rápido.

05 Modelos de Predicción

Ajuste de Hiperparametros:

- Loss Function: especificar la métrica usada durante el entrenamiento qué el algoritmo de gradient boosting va a maximizar/minimizar → En el caso de los modelos de clasificación se utiliza logloss.
- Eval Metric: métrica de evaluación a utilizar → En este caso evaluaremos el modelo con AUROC y Accuracy

- Iterations: cantidad máxima de árboles que se construyen para resolver el problema de Machine Learning. → Cuantas más iteraciones más overfitting, hay que contratarlo. (default 1000)
- Learning Rate: usado para determinar el gradient step (determines the step size at each iteration while moving toward a minimum of a loss function). → Por defecto se define automáticamente según las propiedades del dataset. A menor sea menor overfitting
- Tree Depth: profundidad del árbol. → Se recomienda entre 6 y 12.
- L2 Regularization: fuerza qué remueve un porcentaje pequeño de peso en cada iteración → probar cual es el mejor.

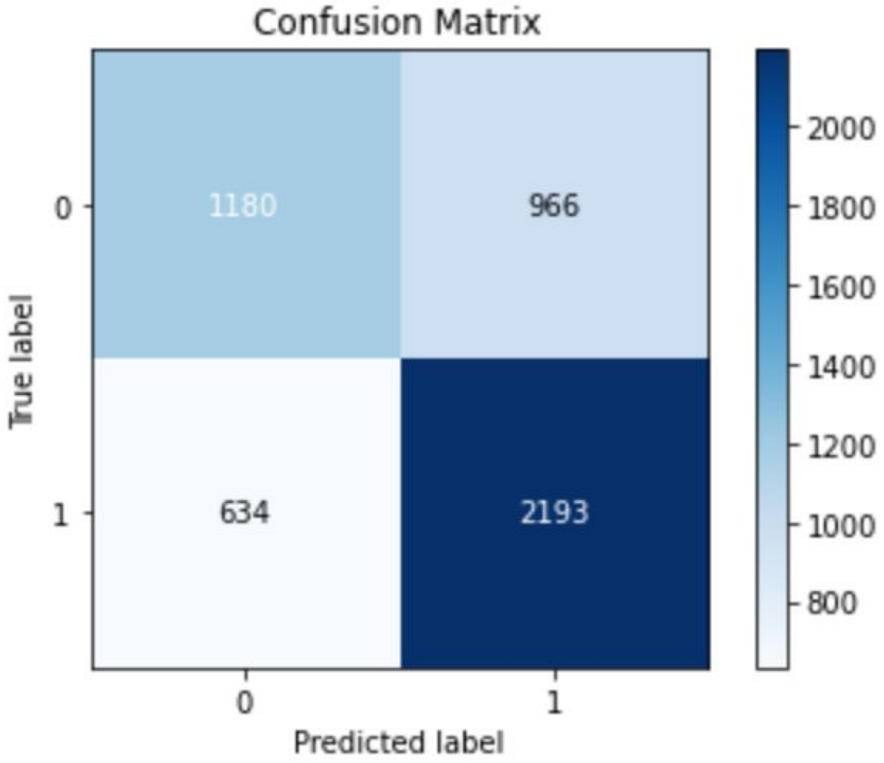
Fuente: https://catboost.ai/en/docs/concepts/parameter-tuning

06 Fitting

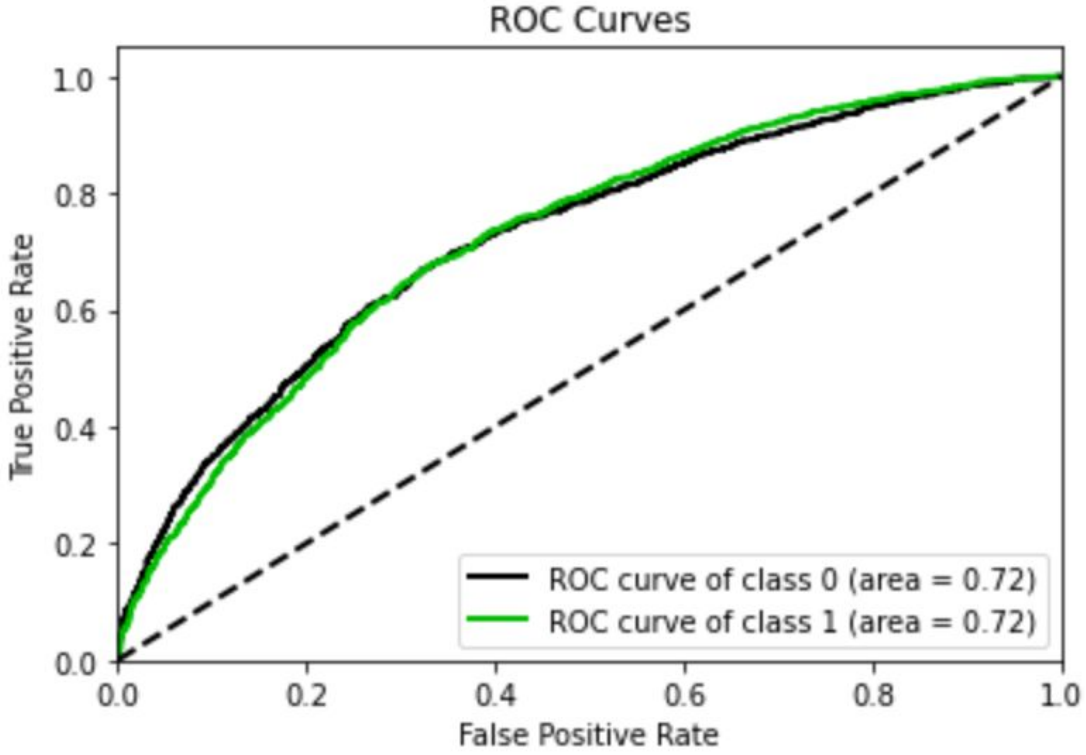
model = CatBoostClassifier(loss_function="Logloss", depth = 10, l2_leaf_reg = 3, iterations = 400, learning_rate = 0.03)

- **AUROC:** 0.663
- **Accuracy:** 0.68

06 Fitting


Confusion Matrix
1180
966
634
2193
0
1
True label
0
1
Predicted label
2000
1800
1600
1400
1200
1000
800

06 Fitting


ROC Curves
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class 0 (area = 0.72)
ROC curve of class 1 (area = 0.72)

ITBA
Instituto Tecnológicc
de Buenos Aires

Gracias!