Aprendizaje Automático

Práctica 2

PREDICCIÓN DEL ABANDONO (BURNOUT) DE EMPLEADOS

2,5 puntos

INTRODUCCIÓN

El objetivo de este segundo trabajo es la construcción de modelos con diversos preprocesos. El tema es el desgaste de empleados: una empresa está preocupada por el nivel de desgaste de los empleados y le gustaría crear un modelo que prediga si es probable que los empleados abandonen la empresa, usando un conjunto de datos recopilados por el departamento de recursos humanos.

CONSIDERACIONES GENERALES

1. Para realizar la práctica, los estudiantes emplearán un repositorio de código en GitHub. Para ello, cada grupo debe crear un repositorio de código privado y agregar como «colaborador» al profesor de prácticas (que indicará a los estudiantes su nombre de usuario en GitHub). Durante la primera semana, el grupo hará llegar al profesor de prácticas el enlace al repositorio de GitHub donde se harán los *commits* (debe haber un único repositorio por grupo). Se espera que cada grupo haga un *commit* semanal del código de la práctica. Esta parte de la práctica se valorará con 0.25 puntos. Además, también habrá que entregar el cuaderno (*notebook*) final a través de Aula Global.
2. Los resultados deben ser reproducibles. Por lo tanto, hay que fijar la semilla de números aleatorios en los lugares adecuados. Se usará como semilla el NIA de uno de los miembros del grupo o bien el número del grupo de prácticas.
3. Para cada grupo, se proporciona un conjunto de datos en formato pickle: *attrition_available_**xx**.pkl* (**xx** representa el número de grupo).

DESARROLLO DE LA PRÁCTICA

1. (**0.25 puntos**) Preparar un repositorio privado en GitHub para poder hacer los *commits* semanales de lo realizado en la práctica cada semana. Haciendo al menos un *commit* cada semana se obtienen 0.25 puntos. Se recomienda que el nombre del repositorio sea vuestro número de grupo de prácticas seguido con el literal "Practica2". Por ejemplo, si sois el grupo 13 de prácticas, el repositorio se llamará "Grupo13-Practica2". Enviar el enlace del repositorio al profesor de prácticas por e-mail.
2. **(0,15 puntos) Hacer un EDA muy simplificado**: cuántas instancias / cuantos atributos y de qué tipo (numéricos, ordinales, categóricos); columnas constantes o innecesarias; que proporción de *missing values* por atributo; tipo de problema: (clasificación o regresión); ¿es desbalanceado?
3. En esta práctica la evaluación será más sencilla que en la primera. Simplemente dividiremos los datos en un conjunto de train para entrenar y ajustar hiper-

parámetros, y un conjunto de test en el que evaluaremos las distintas posibilidades que se probarán en la práctica. Hay que recordar que En problemas de clasificación desbalanceados hay que usar particiones estratificadas y métricas adecuadas (balanced_accuracy, f1, matriz de confusión). También es conveniente que los métodos de construcción de modelos traten el desbalanceo, usando por ejemplo el parámetro *class_weight="balanced"*.

4. **(1.3 puntos) Construcción de modelos**: para esta práctica usaremos **LogisticRegression** como método base (sin ajustar hiper-parámetros) y Boosting como método avanzado (ajustando hiper-parámetros), a elegir. Es importante realizar los preprocesos que los datos necesiten, usando preferentemente **pipelines**. Como método de boosting, se puede elegir uno de entre los métodos de **boosting** disponibles en scikit-learn. **Si además** se usa uno de entre las **librerías externas** xgboost, lightgbm o catboost, se pueden sacar +0.35 puntos adicionales.
5. **(0.8 puntos) Usando algún método de selección de atributos de tipo filter (SelectKBest) de entre los disponibles en sklearn (f_classif, mutual_info_classif o chi2),** comprobad si se pueden mejorar los resultados del apartado anterior y extraer conclusiones sobre qué atributos son más importantes, al menos de acuerdo a estos métodos.

QUÉ ENTREGAR

- Código en un notebook. Es necesario que a lo largo de la práctica se vayan extrayendo conclusiones, y al final de la práctica, hay que hacer un resumen de todos los resultados obtenidos, usando tablas y/o gráficos.
- El archivo conteniendo el mejor modelo obtenido (llamado «modelo_final.pkl»).
- Se recuerda que además de la entrega final, cada semana hay que hacer al menos un *commit* en el GitHub privado de cada grupo (0.25 puntos).

Column Name	Column Description	Data Type
hrs	The number of hours worked by the employee	float64
absences	The number of absences taken by the employee	float64
JobInvolvement	The level of involvement the employee has in their job	float64
PerformanceRating	The employee's performance rating	float64
EnvironmentSatisfaction	The level of satisfaction the employee has with their work environment	float64
JobSatisfaction	The level of satisfaction the employee has with their job	float64
WorkLifeBalance	The balance between work and personal life for the employee	float64
Age	The age of the employee	float64
Attrition	Whether the employee has left the company or not	object
BusinessTravel	The frequency of the employee's business travel	object
Department	The department the employee works in	object
DistanceFromHome	The distance from the employee's home to their workplace	float64
Education	The highest level of education attained by the employee	int64
EducationField	The field of study the employee specialized in	object
EmployeeCount	The number of employees in the company	float64
EmployeeID	A unique identifier for each employee	int64
Gender	The gender of the employee	object
JobLevel	The employee's job level in the company hierarchy	float64
JobRole	The specific role the employee has in their department	object
MaritalStatus	The employee's marital status	object
MonthlyIncome	The employee's monthly income	float64
NumCompaniesWorked	The number of companies the employee has worked for before joining the current company	float64
Over18	Whether the employee is over 18 years old (presumably all employees are)	object
PercentSalaryHike	The percentage of salary increase the employee received in their last salary hike	float64
StandardHours	The standard number of working hours in the company	float64
StockOptionLevel	The level of stock option the employee has	float64
TotalWorkingYears	The total number of years the employee has worked	float64

TrainingTimesLastYear	The number of times the employee received training in the last year	float64
YearsAtCompany	The number of years the employee has been with the company	float64
YearsSinceLastPromotion	The number of years since the employee's last promotion	float64
YearsWithCurrManager	The number of years the employee has been with their current manager.	float64