FACULTAD DE INGENIERIA
Universidad de Buenos Aires

U.B.A. FACULTAD DE INGENIERÍA

Organización de Datos

Primer cuatrimestre 2021

Curso: 02

Trabajo Práctico N2

Machine Learning

Alumno	Número de padrón	Email
Izquierdo Osorio, Stephanie	104196	sizquierdo@fi.uba.ar
Rosas, Martin Alejandro	98535	mrosas@fi.uba.ar

1. Introducción

Luego de la presentación del informe y el baseline FiuFip se profundizó sobre la campaña de recaudación. Gracias al éxito logrado en la primera campaña se realizaron pruebas con distintas técnicas de inteligencia artificial para resolver este problema de clasificación.

2. Archivos

A continuación nombraremos los archivos entregados:

2.1. Notebooks

1. DecisionTree
2. RandomForest
3. NaiveBayes
4. KNN
5. LogisticRegresion
6. Adaboost
7. GradientBoost
8. Voting

2.2. Archivos Auxiliares

1. Preprocessing: Archivo .py donde se encuentran los distintos pre procesamientos para aplicar al dataset antes de entrenar los modelos.

2. Gráficos y escritura: Archivo .py donde se encuentran los códigos para los distintos ploteos de gráficos y la escritura de la predicción.

3. Requirements: Archivo de texto plano donde se especifican los requerimientos para poder ejecutar los modelos.

4. PrediccionesHoldout : Carpeta donde se encuentran las predicciones de los modelos respecto al holdout.

3. Tablas

3.1. Tabla N 1: Pre Procesamientos

La siguiente tabla lista todos los pre procesamientos que fueron aplicados:

Preprocesamiento	Explicación	Nombre de la función

Preproceso básico	Aplica feature engineering con lo visto en el tp1	prepararDatosExistentes
Encodeo Ordenado	Encodea numéricamente respetando el orden previsto a la problemática	encodingsorted
Encodeo one hot	Realiza one hot encoding a las columnas que se le especifique	encodeonehot
Normalización	Normaliza las columnas que se le especifique	normalizarcolumnas

3.2. Tabla N 2: Métricas de los modelos

En la siguiente tabla se presentan los modelos, el preprocesamiento final y las métricas obtenidas:

Modelo	Preprocesamiento	AUC-ROC	Accuracy	Precisión	Recall	F1-Score
Decision Tree	EncodeOneHot	0.910	0,864	0,617	0,761	0,681
Random Forest	OHE y Numeric	0.919	0,869	0,618	0.784	0.691
Naive Bayes	OHE y Numeric	0.896	0.791	0.825	0.539	0.652
KNN	OHE, Numeric, Normalización	0,888	0,847	0,580	0,719	0,642
Logistic Regression	OHE, Numeric	0,910	0,858	0,609	0,743	0,669
Adaboost	OHE, Numeric	0,922	0,868	0,626	0,774	0,692
Gradient Boost	OHE, Numeric	0,924	0,872	0,611	0,803	0,694
Voting	OHE, Numeric	0,920	0,865	0,578	0,799	0,671

4. Conclusiones

En el trabajo práctico probamos 8 clasificadores distintos, cada uno con distintos pre procesamientos e hiperparametros. Finalmente el modelo que recomendamos es el que nos dio los mejores resultados utilizando el área bajo curva de ROC como métrica: Grandient Boost. Los bootsters fueron los modelos que mejores números arrojaron. Respecto a la función baseline del TP1, la cual tuvo un accuracy de 81.4%, notamos que en varios de nuestros modelos pudimos superar este valor.

En el caso de buscar la menor cantidad de falsos positivos, tendríamos que elegir el modelo que tenga mayor precisión. En ese caso el mejor sería Naive Bayes. De esa manera FIUFIP estaría yendo por el camino que más le evite asumir como alto valor adquisitivo a gente que no lo sea. En cambio, si la idea es atrapar a todos los casos de alto valor adquisitivo que podamos, sin importar que acusemos a muchos sin realmente serlo, deberíamos buscar al modelo con mayor recall. Este va a ser Gradient Boost.