

PRÁCTICA Nº 1

KDD con BigML

- Módulo: Sistemas de BD
- Nombre y apellidos: Alvaro Lucio-Villegas de Cea



Índice

Enunciado:	3
Proceso Análisis con BigML	4
Carga de Datos	4
División de Dataset	5
Creación Modelo	6
Predicción de modelo	7
Creación Ensemble	9
Predicción de Ensemble	12
Batch Prediction	13
Evaluate	15
Preguntas:	16

Enunciado:

Realiza la práctica guiada que encontrarás en el siguiente enlace:

<https://cleverdata.io/machine-learning-prediccion-basada-datos-bigml/>

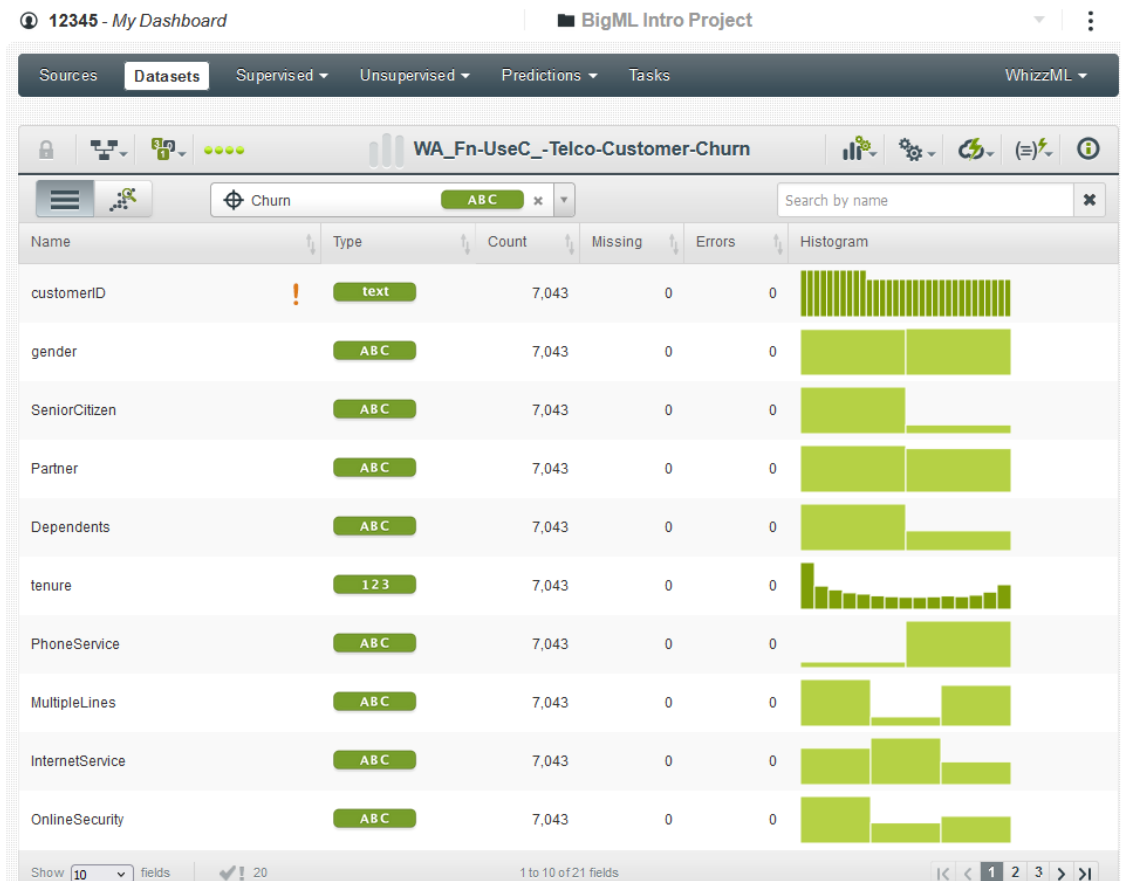
Redacta una memoria con los pasos realizados. Incluye respuestas a las siguientes cuestiones:

- Responde a las preguntas que el propio artículo va planteando.
- Asocia el proceso realizado a las diferentes etapas de las que consta un proceso KDD. Si alguna etapa no se ha realizado, indica por qué motivo. Si alguna etapa se ha realizado de forma "automática" indícalo.
- ¿Los datos de partida se pueden considerar estructurados? ¿y etiquetados? ¿es lo normal?. Si los datos no estuvieran estructurados ni etiquetados, ¿qué habiéramos tenido que hacer?
- Realiza la mejora de la predicción del primer artículo que se propone mediante el uso de [Ensembles](#).

Proceso Análisis con BigML

Carga de Datos

Cargamos los datos en la aplicación.



División de Dataset

Para poder realizar el análisis deberemos de dividir el dataset en 2. Usaremos la herramienta que nos proporciona la aplicación.

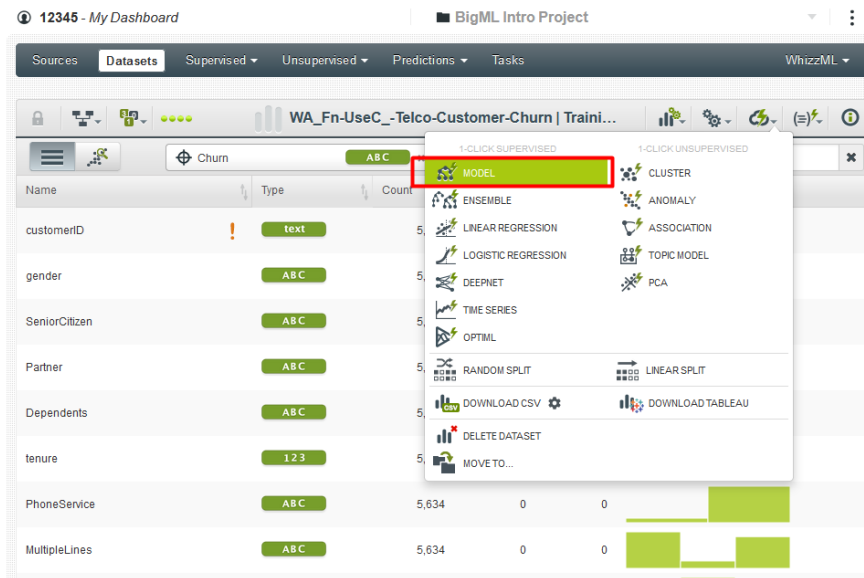
The screenshot shows the WhizzML interface with the 'WA_Fn-UseC_-Telco-Customer-Churn' dataset selected. A dropdown menu is open, showing options for 'SAMPLE AND FILTER DATASET' and 'TRANSFORM DATASET'. The 'TRAINING | TEST SPLIT' option is highlighted in red. The dataset table below shows various features like customerID, gender, SeniorCitizen, Partner, Dependents, tenure, PhoneService, MultipleLines, InternetService, and OnlineSecurity, along with their types and distributions.

En este caso lo dividiremos en 80% de los datos para entrenamiento y un 20% para test.

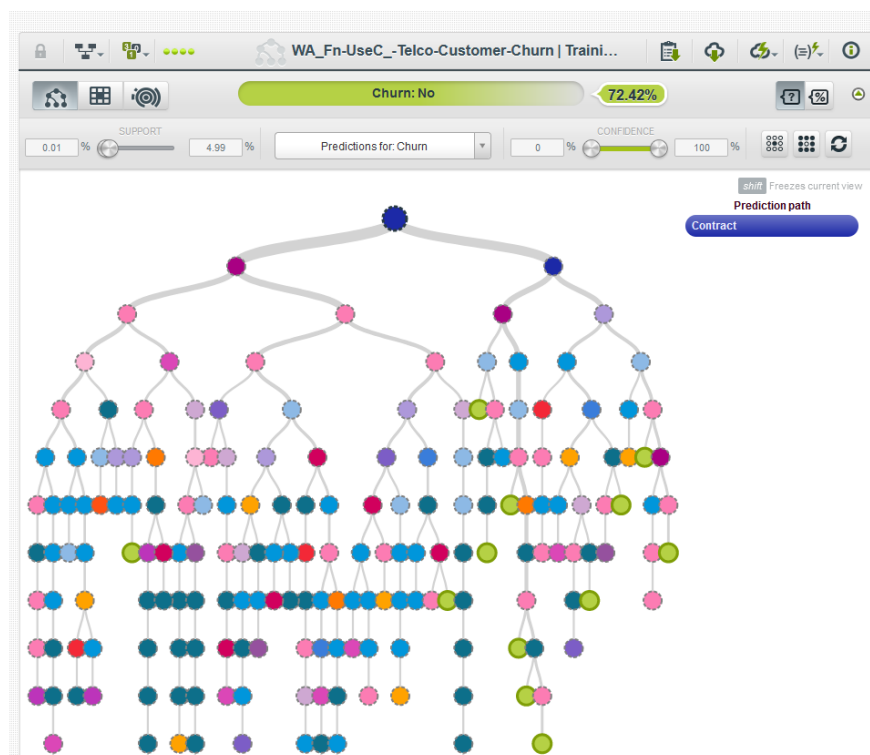
The screenshot shows the 'SPLIT DATASET CONFIGURATION' window. It displays a slider for 'Training' at 80% and 'Test' at 20%. The 'Seed' field is set to 0. The 'Linear split' option is selected. The 'Training dataset name' and 'Test dataset name' fields are both set to 'WA_Fn-UseC_-Telco-Customer-Churn'. A 'Create Training | Test' button is visible at the bottom right.

Creación Modelo

Una vez divididos usaremos el dataset de entrenamiento para crear el modelo.

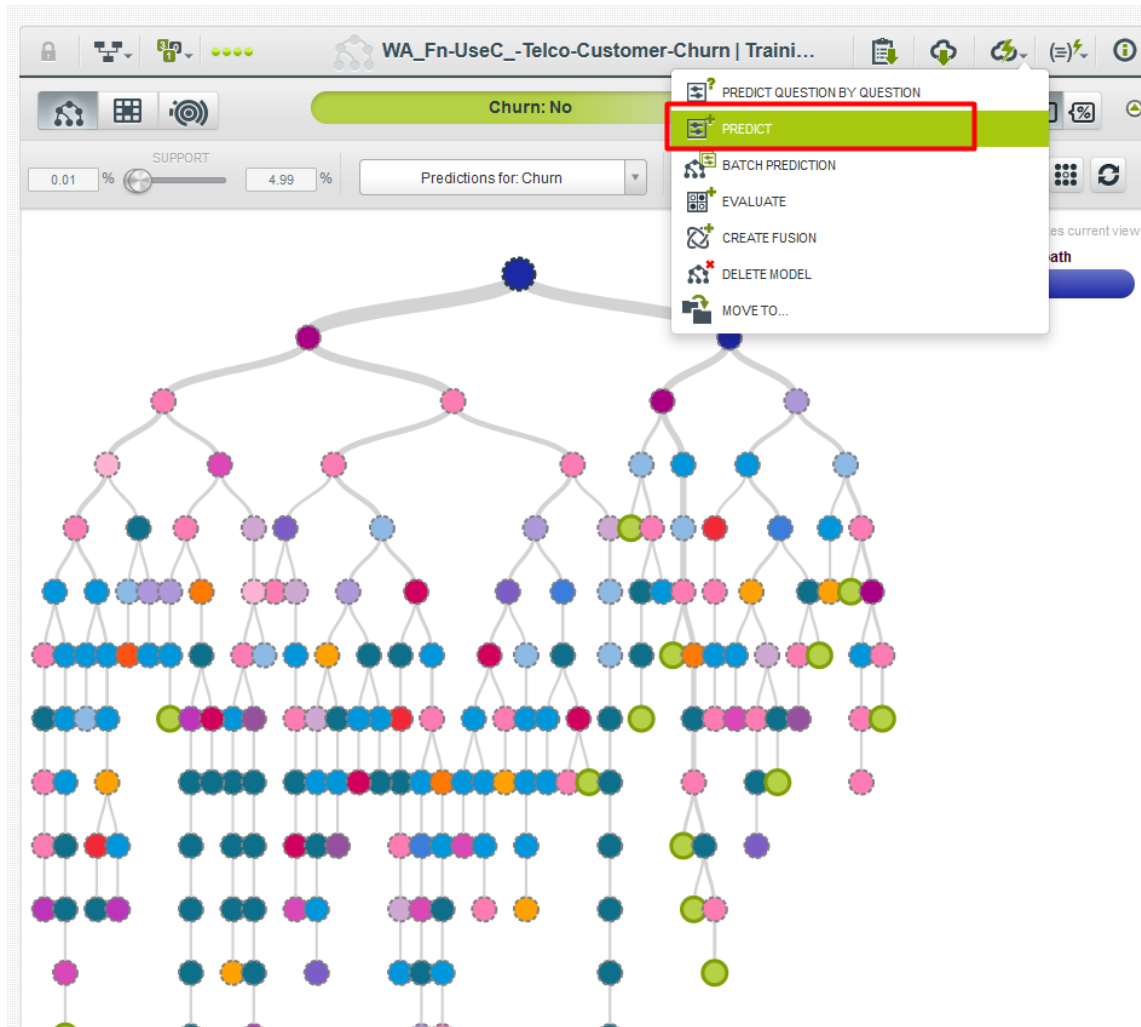


Como podemos observar ya tenemos el árbol creado y por ahora podemos observar que tenemos un 72% de predicción en el modelo.

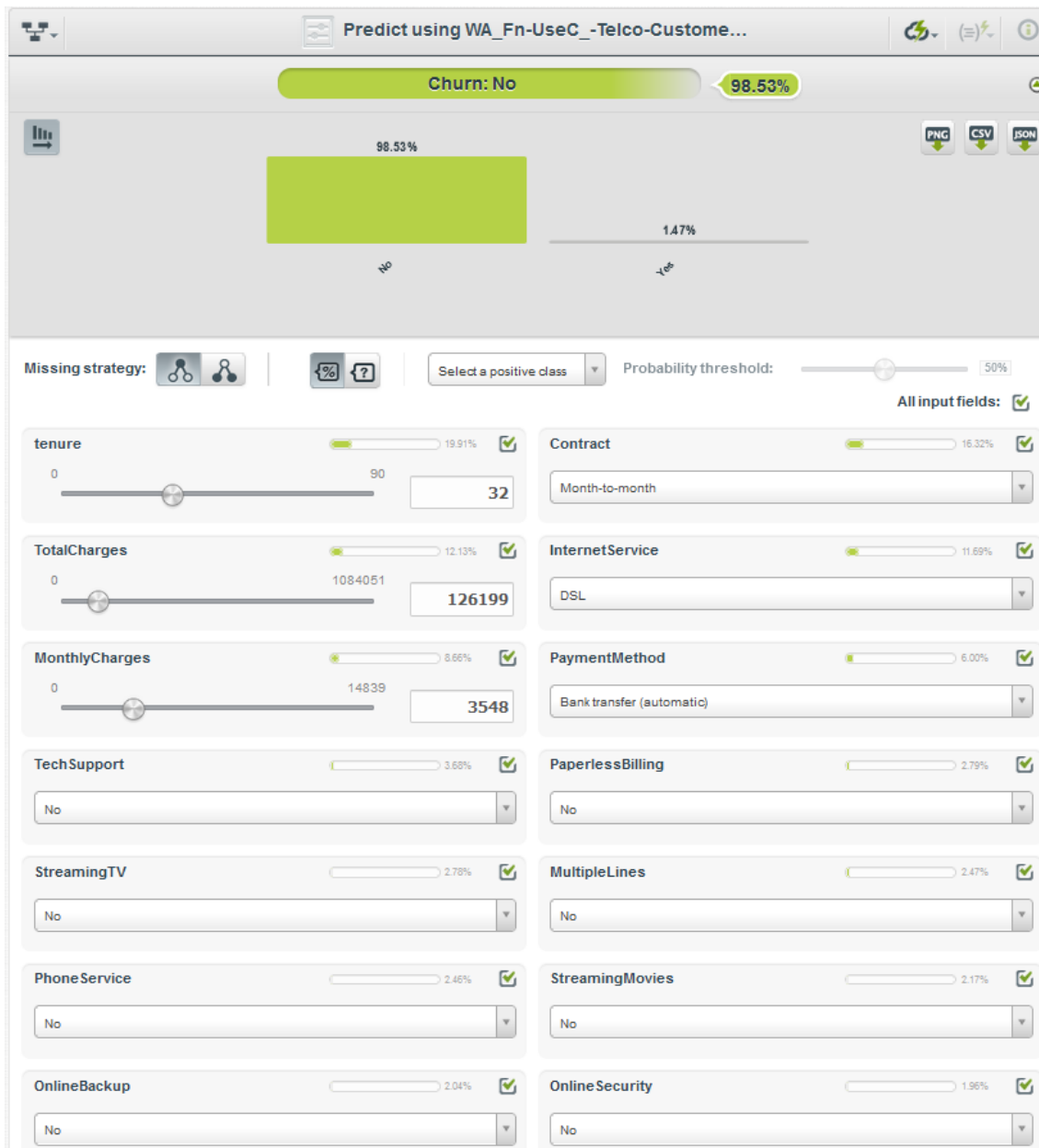


Predicción de modelo

Vamos a usar la herramienta de la aplicación para que nos visualice el valor de la predicción



Aquí podemos observar los distintos valores del dataset con los que podemos ir viendo qué valores son los más influyentes para la predicción final.



Creación Ensemble

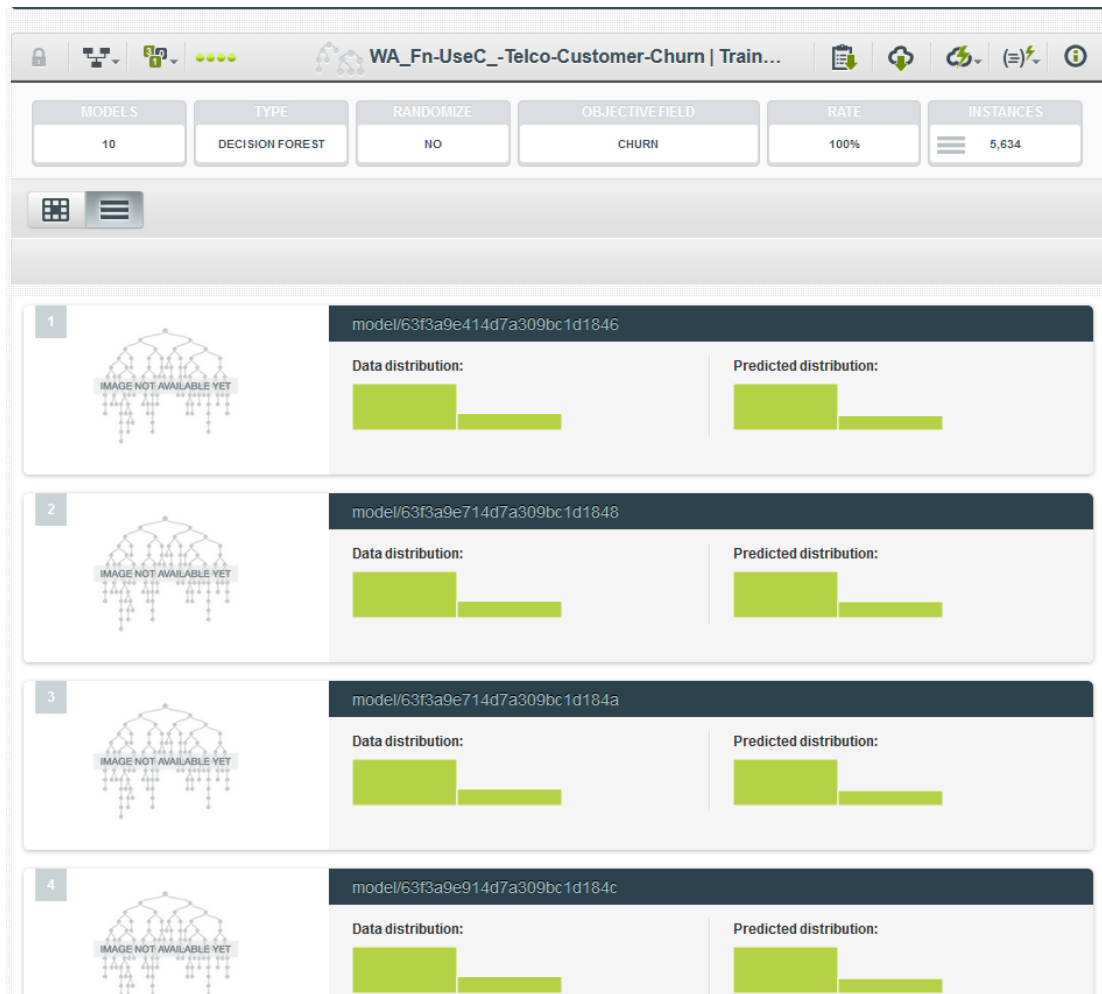
Para poder obtener mejores predicciones podemos crear un ensemble que creara más árboles y nos aportará más precisión sobre los resultados

The screenshot shows the WhizzML dashboard for a project named 'WA_Fn-UseC_-Telco-Customer-Churn'. The 'Datasets' tab is active, displaying a table with columns: Name, Type, Count, Missing, Errors, and Histogram. The 'Churn' variable is selected. A dropdown menu is open, showing '1-CLICK SUPERVISED' options. The 'ENSEMBLE' option is highlighted with a red box. Other options include MODEL, LINEAR REGRESSION, LOGISTIC REGRESSION, DEEPMET, TIME SERIES, OPTIML, CLUSTER, ANOMALY, ASSOCIATION, TOPIC MODEL, and PCA. Below the menu, there are buttons for 'RANDOM SPLIT', 'LINEAR SPLIT', 'DOWNLOAD CSV', 'DELETE DATASET', and 'MOVETO...'.

Seleccionamos la optimización automática y creamos el ensemble.

The screenshot shows the 'ENSEMBLE CONFIGURATION' window in WhizzML. The 'Objective field' is set to 'Churn'. The 'Automatic optimization' checkbox is checked. The 'Type' is set to 'Decision Forest'. The 'Number of models' is set to 10, and the 'Number of iterations' is set to 64. The 'Ensemble name' is 'WA_Fn-UseC_-Telco-Customer-Churn | Training (80%)'. There are buttons for 'Reset' and 'Create ensemble'. Below the configuration, there is a table with columns: Name, Type, Count, Missing, Errors, and Histogram. The table lists variables: customerID, gender, SeniorCitizen, Partner, Dependents, tenure, and PhoneService. Each variable has a corresponding histogram.

Podemos observar los distintos árboles que se han creado.



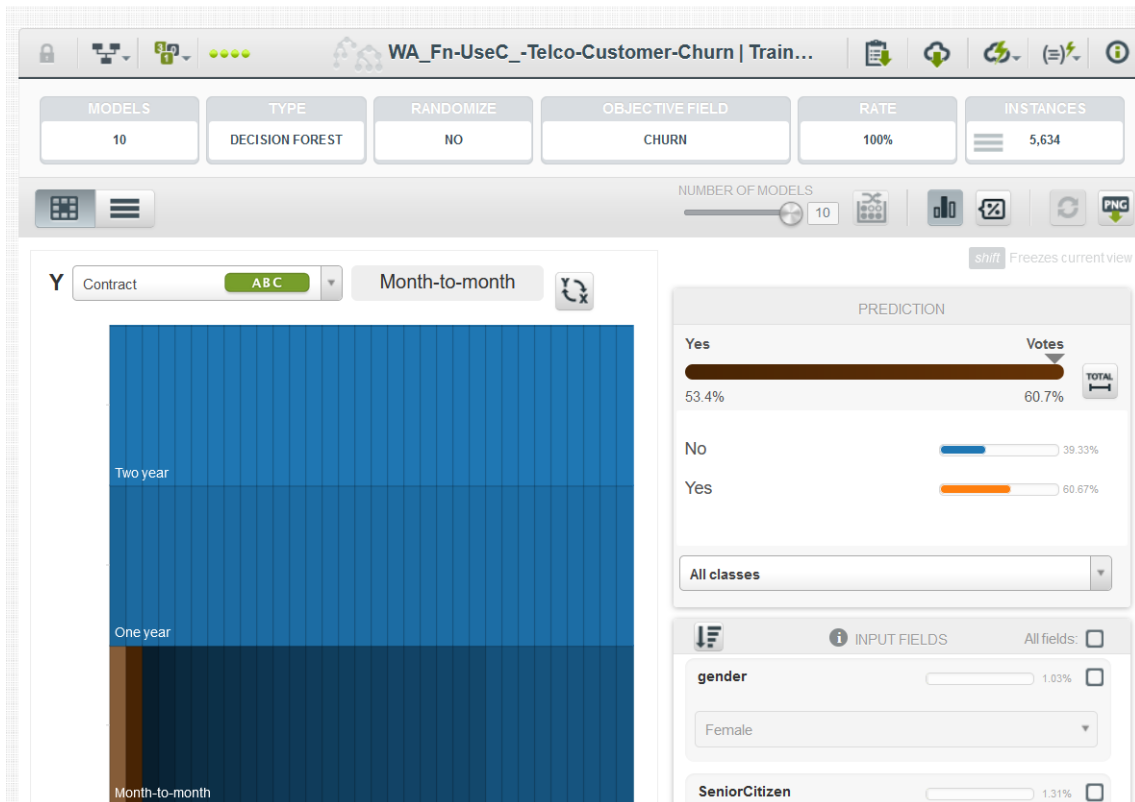
Configuramos la decisión para que la predicción sea positiva.

Según la configuración anterior podemos ver el resultado de predicción del resto de los árboles.

Model ID	Prediction	Probability
model/63f3a9e414d7a309bc1d1846	Yes	90.82%
model/63f3a9e714d7a309bc1d1848	Yes	94.36%
model/63f3a9e714d7a309bc1d184a	No	73.90%
model/63f3a9e914d7a309bc1d184c	Yes	84.34%
model/63f3a9ea14d7a309bc1d184e	Yes	89.62%
model/63f3a9ea14d7a309bc1d1850	Yes	81.15%
model/63f3a9eb14d7a309bc1d1852	No	97.61%
model/63f3a9eb14d7a309bc1d1854	No	74.94%
model/63f3a9ed14d7a309bc1d1856	No	83.63%
model/63f3a9ee14d7a309bc1d1858	No	96.42%

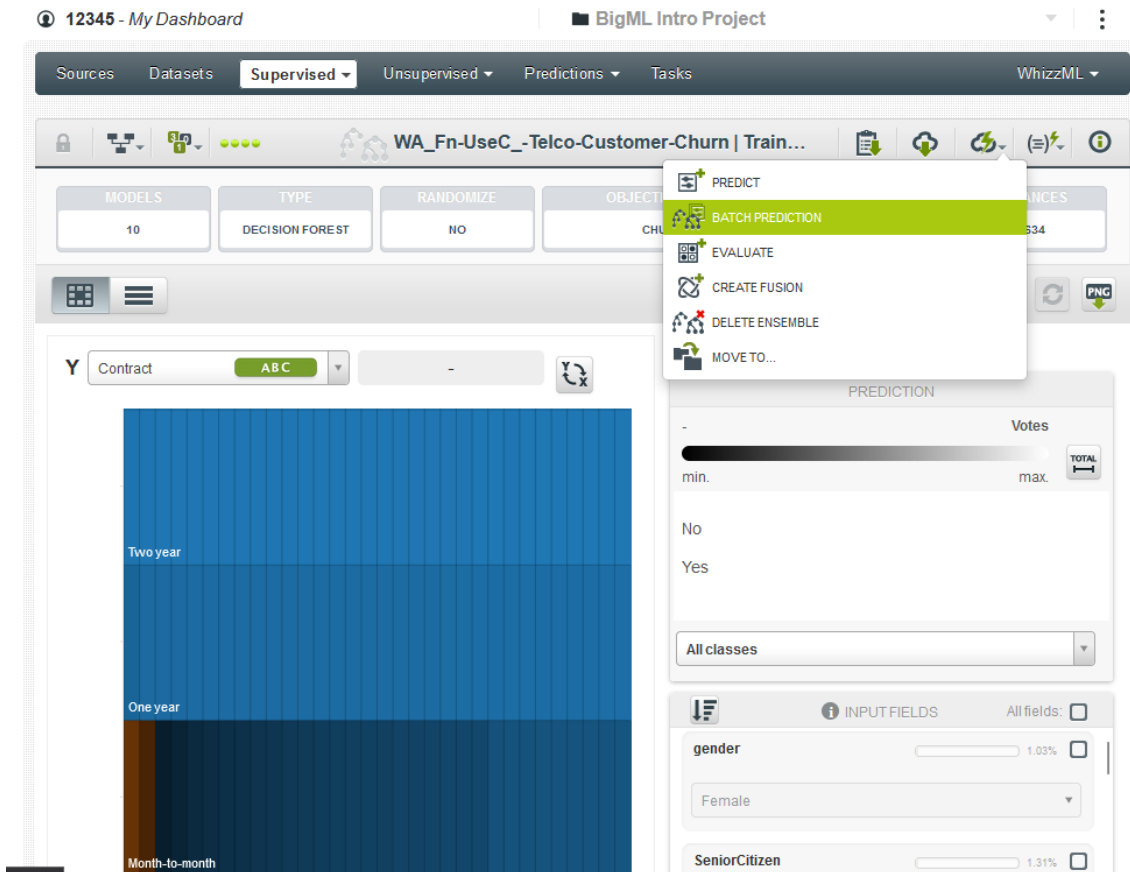
Predicción de Ensemble

Podemos ver una matriz que muestra la predicción dependiendo de los valores de las columnas.



Batch Prediction

Ahora haremos la comprobación del modelo con el dataset de test del principio, para ello usaremos la herramienta de “Batch Prediction” sobre el ensemble.



The screenshot shows the BigML web interface. At the top, there's a navigation bar with 'Sources', 'Datasets', 'Supervised', 'Unsupervised', 'Predictions', and 'Tasks'. Below this, a project titled 'WA_Fn-UseC_-Telco-Customer-Churn | Train...' is open. A dropdown menu is open, showing options: 'PREDICT', 'BATCH PREDICTION' (highlighted), 'EVALUATE', 'CREATE FUSION', 'DELETE ENSEMBLE', and 'MOVE TO...'. The main area shows a decision tree visualization with nodes labeled 'Two year', 'One year', and 'Month-to-month'. On the right, there's a 'PREDICTION' section with a 'Votes' slider and a 'TOTAL' button. Below that, the 'INPUT FIELDS' section shows 'gender' (Female) and 'SeniorCitizen' (1.31%).

Seleccionamos a la izquierda el ensemble y a la derecha el dataset de test. Y le daríamos a "Predict".

New Batch Prediction

WA_Fn-UseC_-Telco-Customer-Churn | Training (80%) x

Churn Mon, 20 Feb 2023 17:11:54

763.6 KB size 20 fields 5,634 instances 10 models

Description:

WA_Fn-UseC_-Telco-Customer-Churn | Test (20%) x

Mon, 20 Feb 2023 16:55:22

191.0 KB size 21 fields 1,409 instances

Description:

Configure

Preview of the prediction file (using the type of each field)

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	Phone Service	MultipleLines	InternetService	OnlineSecurity	O
text	ABC	ABC	ABC	ABC	123	ABC	ABC	ABC	ABC	AI
text	ABC	ABC	ABC	ABC	123	ABC	ABC	ABC	ABC	AI
text	ABC	ABC	ABC	ABC	123	ABC	ABC	ABC	ABC	AI
text	ABC	ABC	ABC	ABC	123	ABC	ABC	ABC	ABC	AI
text	ABC	ABC	ABC	ABC	123	ABC	ABC	ABC	ABC	AI

Prediction name:

WA_Fn-UseC_-Telc...urn | Test (20%) with WA_Fn-UseC_-Telco-C

Reset Predict

Como podemos observar el resultado es al final del daset dos columnas de "Churn" una es el resultado esperado y el otro el resultado real. Y como podemos observar es correcta la predicción del modelo.

WA_Fn-UseC_-Telc...urn | Test (20%) with WA_Fn-Use...

Configuration

MISSING STRATEGY: Last prediction

OPERATING KIND: Probability

DEFAULT NUMERIC: -

Output preview

amingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn	Churn
No	No	One year	No	Bank transfer (automatic)	423	184075	No	No
No	No	Month-to-month	Yes	Credit card (automatic)	891	19494	No	No
No	No	Month-to-month	No	Mailed check	2975	3019	No	No
Yes	Yes	Month-to-month	Yes	Electronic check	1048	304605	Yes	Yes
Yes	Yes	One year	No	Credit card (automatic)	10035	56811	No	No

Download batch prediction

Output dataset

Evaluate

12345 - My Dashboard

BigML Intro Project

Sources Datasets **Supervised** Unsupervised Predictions Tasks WhizzML

New Evaluation

WA_Fn-UseC_-Telco-Customer-Churn | Training (80... *

Churn Mon, 20 Feb 2023 17:11:54

763.6 KB size 20 fields 5,634 instances 10 models

Description:

Configure

WA_Fn-UseC_-Telco-Customer-Churn | Test (20%) *

Mon, 20 Feb 2023 16:55:22

191.0 KB size 21 fields 1,409 instances

Description:

Configure

Evaluation name:

WA_Fn-UseC_-Telco-Customer-Churn | Training (80%) vs. WA_Fi

Reset Evaluate

WA_Fn-UseC_-Telco-Customer-Churn | Training (80...

WA_Fn-UseC_-Telco-Customer-Churn | ...

WA_Fn-UseC_-Telco-Customer-Churn | ...

Positive class: Yes

ACTUAL VS. PREDICTED				ACTUAL	RECALL	F	Phi
No	916	112	1,028	89.11%	0.85	0.40	
Yes	201	180	381	47.24%	0.53	0.40	
PREDICTED	1,117	292	1,409				
PRECISION	82.01%	61.64%	71.83% AVG. PRECISION				

The balanced harmonic mean of precision and recall.

$F\text{-measure} = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$

77.79% Accuracy

0.5349 F-measure

61.64% Precision

47.24% Recall

0.3983 Phi coefficient

Preguntas:

Responde a las preguntas que el propio artículo va planteando. Incluye un pantallazo en el que el pronóstico sea positivo y otro negativo.

¿Te acuerdas del fichero con el 20% de los datos? Es hora de usarlo. Hay que subirlo, crear un Dataset y hacer predicciones “Batch Prediction”.

- Si

¿Ya lo has hecho? El modelo que hemos creado, ¿está haciendo predicciones fiables? ¿Qué porcentaje de las predicciones ha acertado? Te adelantamos que este modelo se puede mejorar, pero eso lo explicaremos en próximos artículos.

- 72% en un árbol de decisión.

- Asocia el proceso realizado a las diferentes etapas de las que consta un proceso KDD. Si alguna etapa no se ha realizado, indica por qué motivo. Si alguna etapa se ha realizado de forma "automática" indícalo.

- Selección: Carga del dataset.

- Preproceso: La limpieza es de forma automática.

- Transformación: La limpieza de forma automática

- Data Mining: Lo realiza de forma automática

- Interpretación/Evaluación: Las opciones de predict y evaluate

- El paso de Preprocesado lo hace de forma automática

¿Los datos de partida se pueden considerar estructurados?

- Si

¿y etiquetados?

- Si, ya que poseen de una columna “churn” que dice el resultado obtenido.

¿Es lo normal?

- No, lo habitual es primero cluterizarlos y etiquetarlos para su posterior análisis.

Si los datos no estuvieran estructurados ni etiquetados, ¿qué habríamos tenido que hacer?

- Los datos no estructurados son más difíciles de analizar y requerirán técnicas de análisis de datos más avanzadas, como el procesamiento del lenguaje natural (NLP), el análisis de imágenes, el reconocimiento de voz, la minería de texto, entre otros.
- Realiza la mejora de la predicción del primer artículo que se propone mediante el uso de [Ensembles](#).