

BIG DATA for Business

2.8 Introducción a Dataiku DSS



Conecta Empleo

Contenido desarrollado por
Synergic Partners



Índice del módulo

2.8 INTRODUCCIÓN A DATAIKU DSS

- ¿Qué es Dataiku DSS?
- Características Generales y Técnicas
- Tareas Básicas con DSS



¿Qué es Dataiku DSS?

INTRODUCCIÓN A DATAIKU DSS

¿Qué es Dataiku DSS?

Dataiku = One Product
Data + Technology + People



End-to-End Solution

Based on Open Source

Collaborative

Production-focused

INTRODUCCIÓN A DATAIKU DSS

¿Qué es Dataiku DSS?



data
iku

“Integrated development platform for data professionals to turn raw data into predictions”

Data Science Studio (DSS) es una plataforma software que integra todas la etapas y herramientas necesarias para crear aplicaciones predictivas y de negocio en entornos *Big Data*.

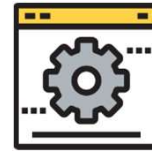


Características Generales y Técnicas

INTRODUCCIÓN A DATAIKU DSS

Características Generales

Automatiza el flujo de trabajo de carga, visualización, y análisis de datos.



Posee una interfaz gráfica interactiva e intuitiva adecuada para los distintos perfiles de un equipo de datos.



Conecta e integra distintas fuentes de datos, como: CSV, bases de datos SQL, MongoDB, HP Vertica, Amazon, Redshift, Hadoop, Spark

Posee librerías propias de machine learning para ajustar modelos de predicción y clustering. Además permite integrar código propio en R, Python, y librerías externas de ML.



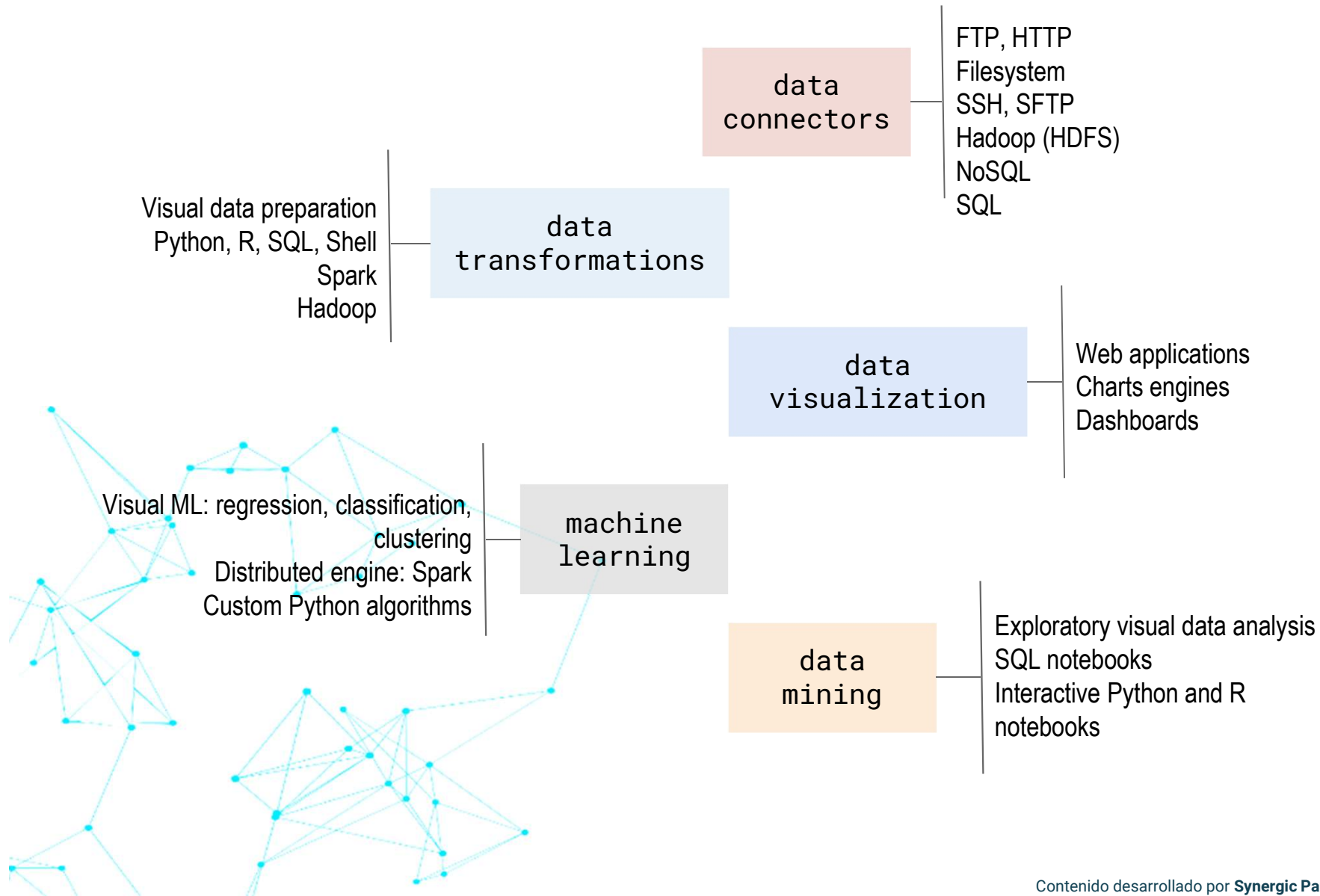
Posee herramientas de monitorización y control de versiones.



Empaqueta el flujo de trabajo para su despliegue en producción.

INTRODUCCIÓN A DATAIKU DSS

Características Técnicas

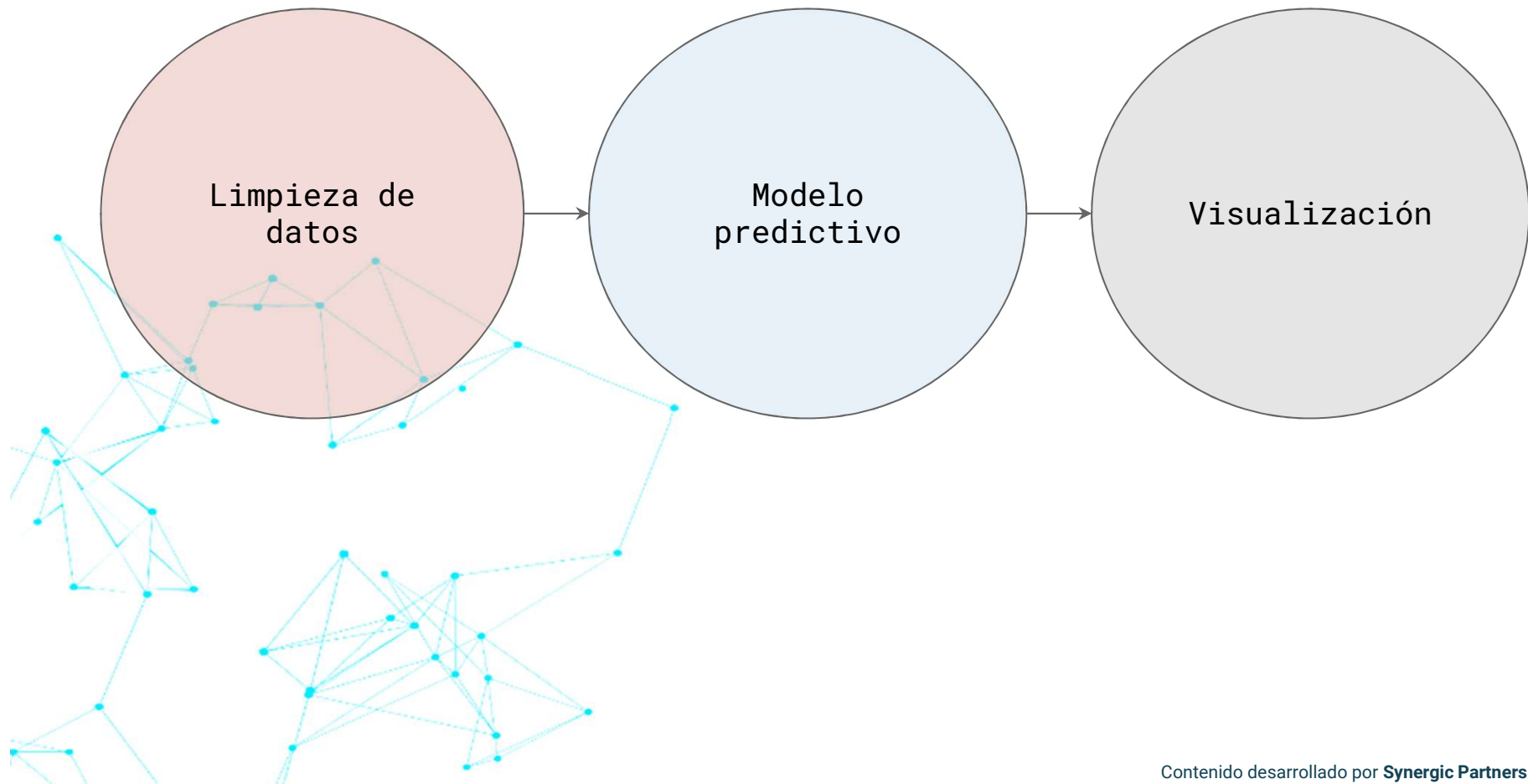




Tareas Básicas con DSS

INTRODUCCIÓN A DATAIKU DSS

Tareas Básicas con DSS

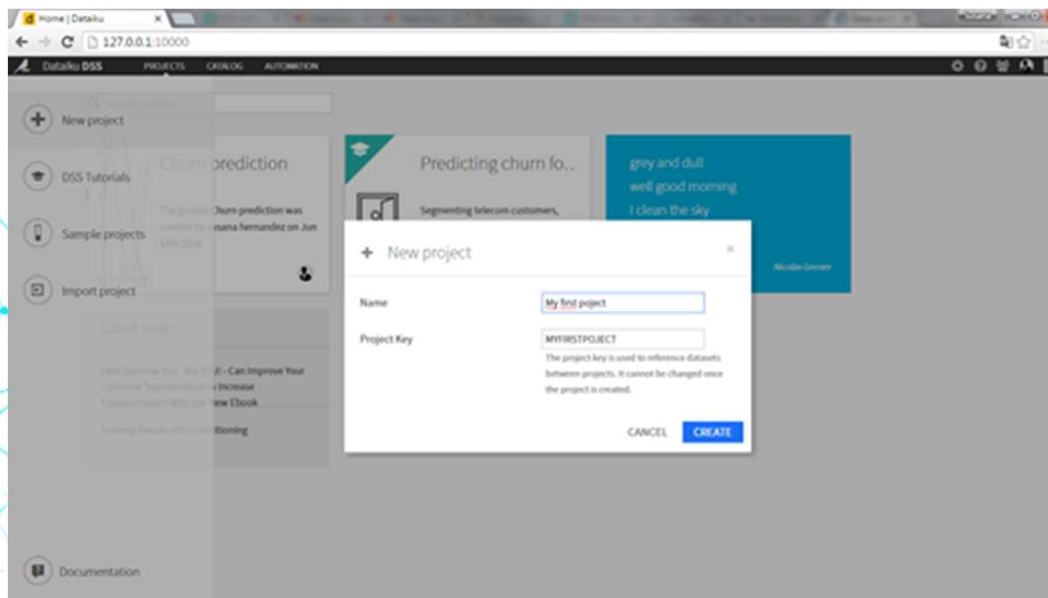


INTRODUCCIÓN A DATAIKU DSS

Tareas Básicas con DSS

LD

1. Crear un proyecto
2. Importar un dataset
3. Vista previa y carga de los datos
4. Limpieza y preparación de los datos



Nuevo proyecto

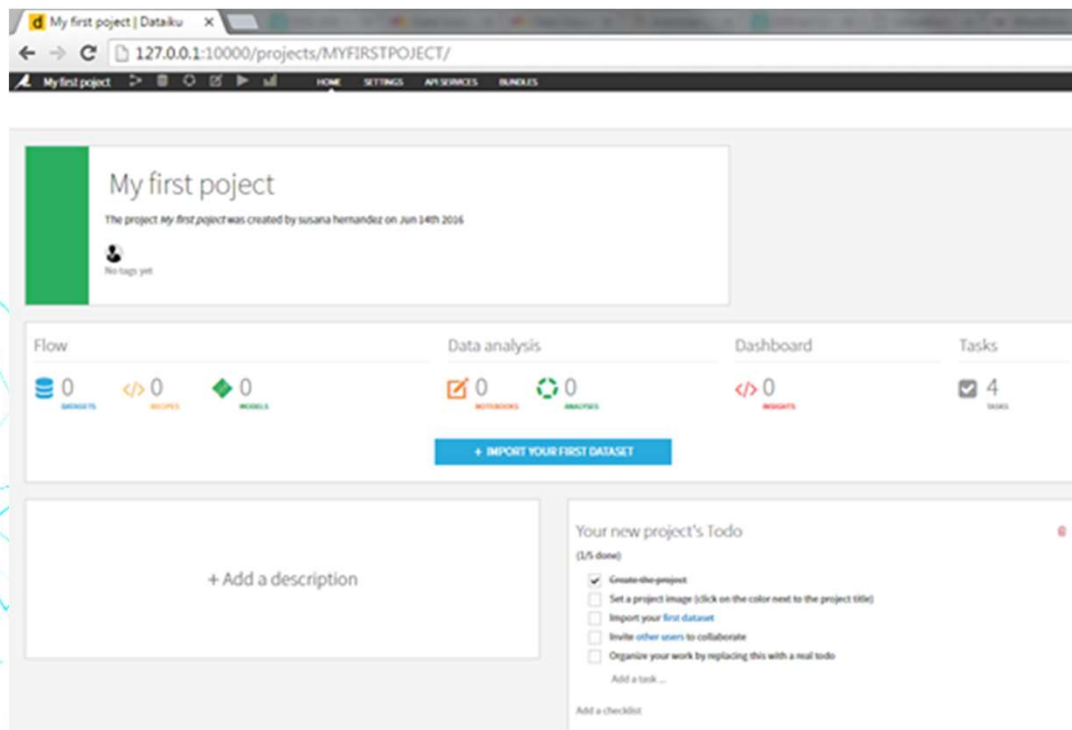
Basta con seleccionar **New Project** en la barra lateral izquierda que aparece al pasar el ratón por esta zona. Podremos identificar nuestro proyecto con un nombre.

INTRODUCCIÓN A DATAIKU DSS

Tareas Básicas con DSS

LD

1. Crear un proyecto
2. **Importar un dataset**
3. Vista previa y carga de los datos
4. Limpieza y preparación de los datos



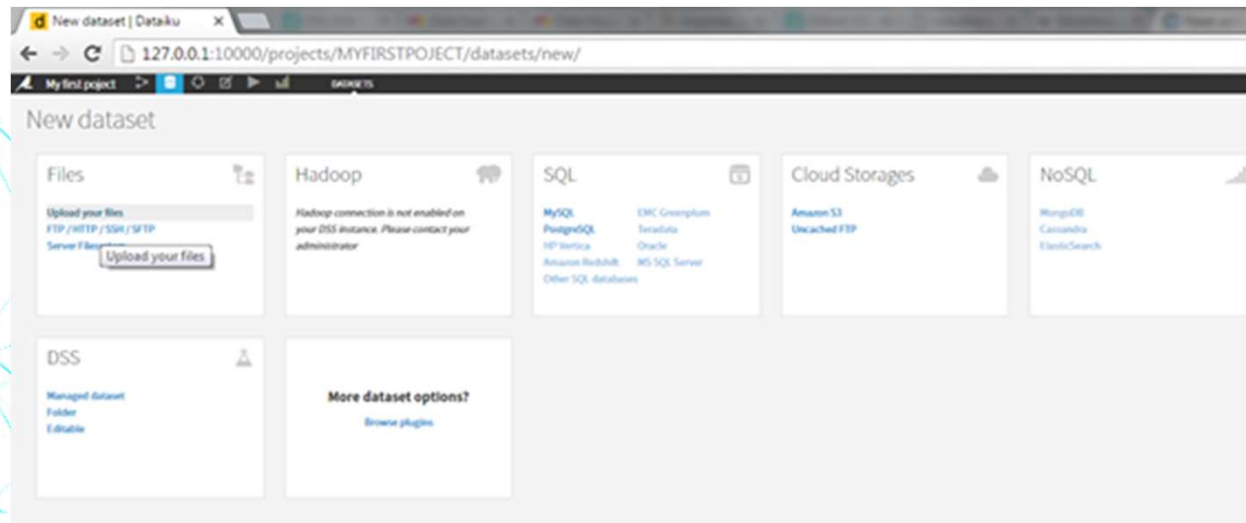
Entramos en el proyecto y añadimos un nuevo **dataset** a partir de la fuente que deseemos.

INTRODUCCIÓN A DATAKU DSS

Tareas Básicas con DSS

LD

1. Crear un proyecto
2. **Importar un dataset**
3. Vista previa y carga de los datos
4. Limpieza y preparación de los datos



Conectores

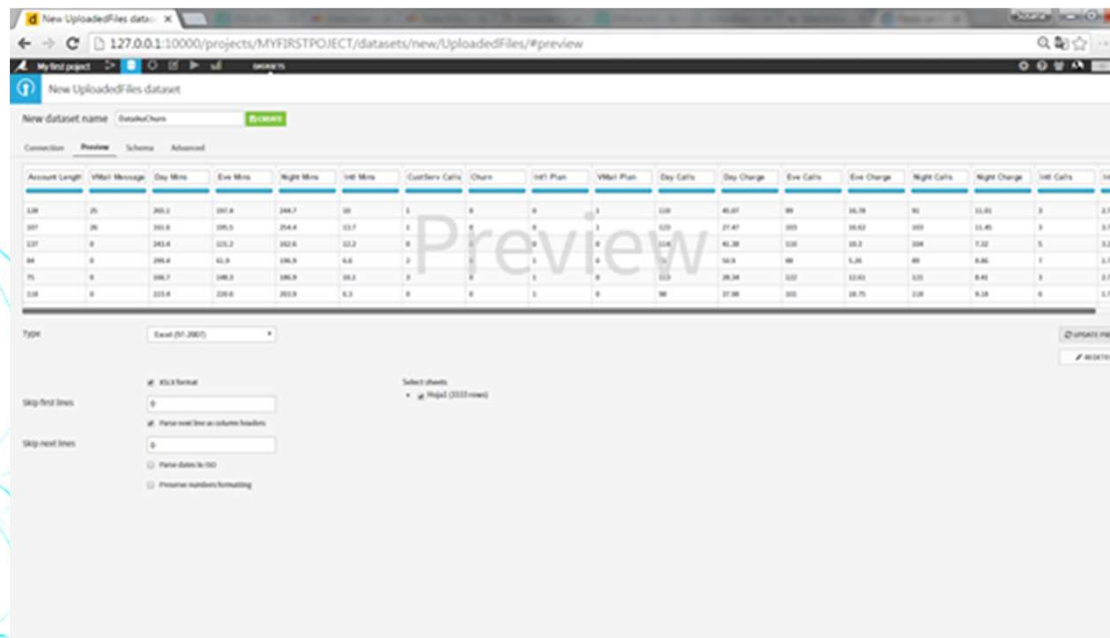
En esta imagen se muestran las distintas fuentes desde las que podemos importar nuestro **dataset**, p.e.: Hadoop, SQL Cloud Storages, NoSQL, etc.

INTRODUCCIÓN A DATAKU DSS

Tareas Básicas con DSS

LD

1. Crear un proyecto
2. Importar un dataset
3. **Vista previa y carga de los datos**
4. Limpieza y preparación de los datos



Account Length	VMail Message	Day Mins	Eve Mins	Night Mins	Int Mins	Customer Calls	Churn	Int Plan	VMail Plan	Day Calls	Day Charge	Eve Calls	Eve Charge	Night Calls	Night Charge	Int Calls	Int Ch
128	25	265.1	267.8	264.7	38	1	0	0	1	128	45.07	89	16.78	81	11.81	9	1.7
107	26	261.8	265.5	264.8	11.7	1	0	0	1	107	27.47	109	16.62	109	11.46	9	1.7
137	0	241.8	115.2	102.6	12.2	0	0	0	0	137	41.38	118	18.3	104	7.17	1	1.19
84	0	266.8	41.9	196.9	6.6	2	0	0	0	84	16.9	88	1.26	88	6.86	1	1.19
75	0	196.7	198.3	198.9	18.1	0	0	0	0	75	28.34	117	11.61	111	8.41	3	1.19
118	0	211.8	218.8	201.9	6.3	0	0	0	0	118	27.08	105	18.71	118	9.18	6	1.7

Una vez subido el fichero podemos hacer una **Preview** de los datos que contiene.

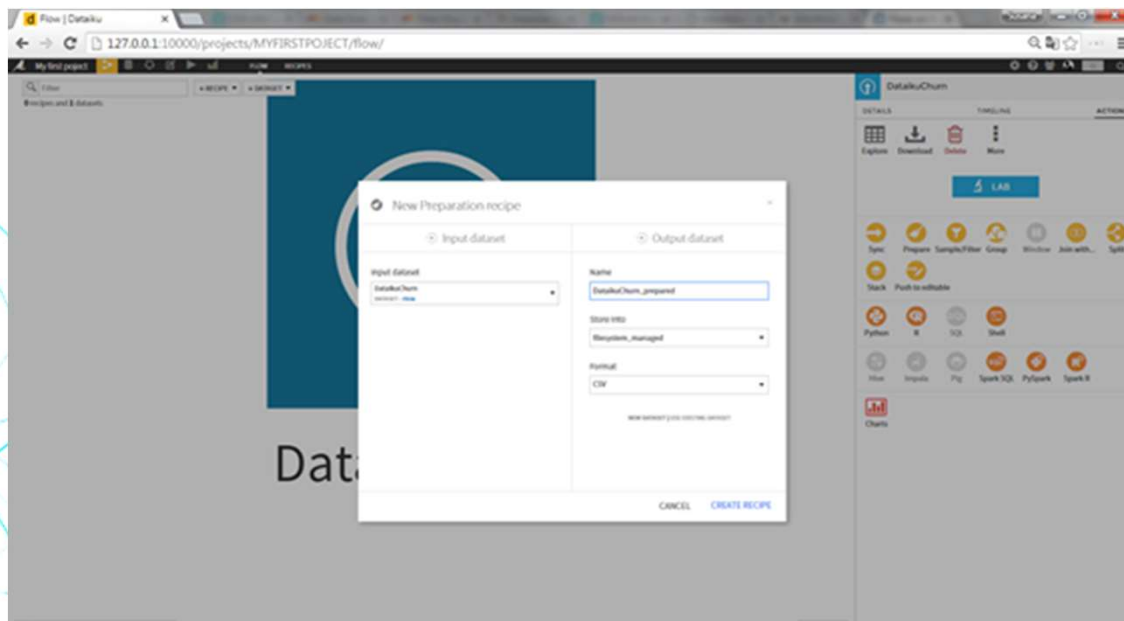
Se pueden realizar las primeras correcciones tales como nombres de columnas, eliminación de líneas vacías al principio del fichero o corrección del tipo de dato detectado.

INTRODUCCIÓN A DATAIKU DSS

Tareas Básicas con DSS

LD

1. Crear un proyecto
2. Importar un dataset
3. Vista previa y carga de los datos
4. Limpieza y preparación de los datos



En el diagrama de flujo, pinchando sobre el **dataset** se nos despliegan las distintas acciones que podemos realizar sobre el mismo.

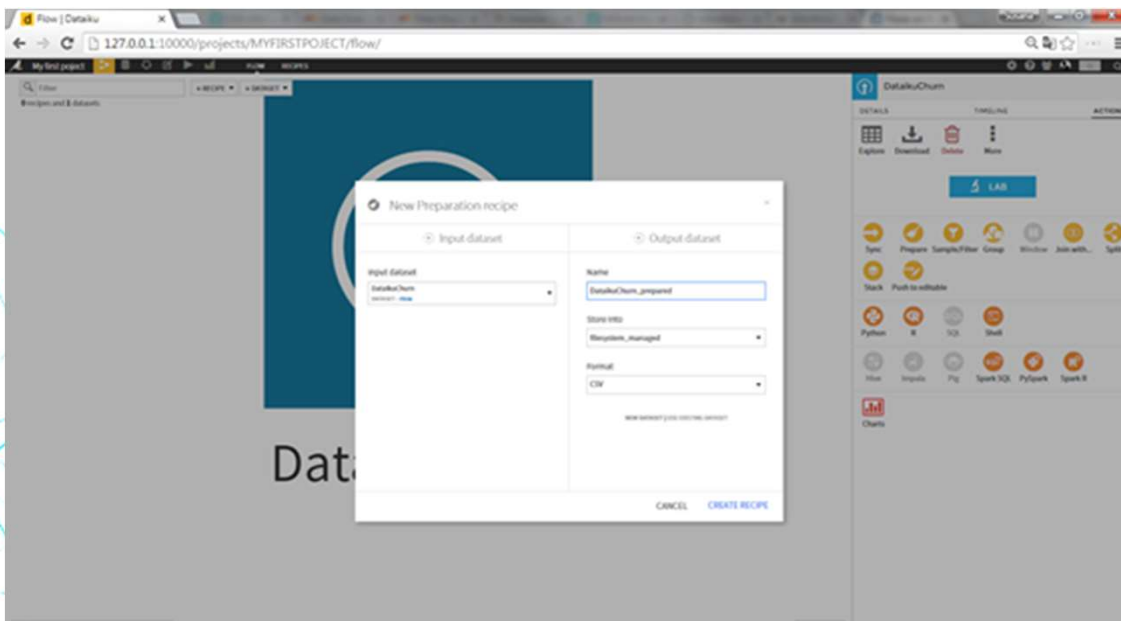
Seleccionaremos la opción de **Prepare**. Se nos generará un nuevo **dataset** con las modificaciones que realicemos.

INTRODUCCIÓN A DATAIKU DSS

Tareas Básicas con DSS

LD

1. Crear un proyecto
2. Importar un dataset
3. Vista previa y carga de los datos
4. **Limpieza y preparación de los datos**



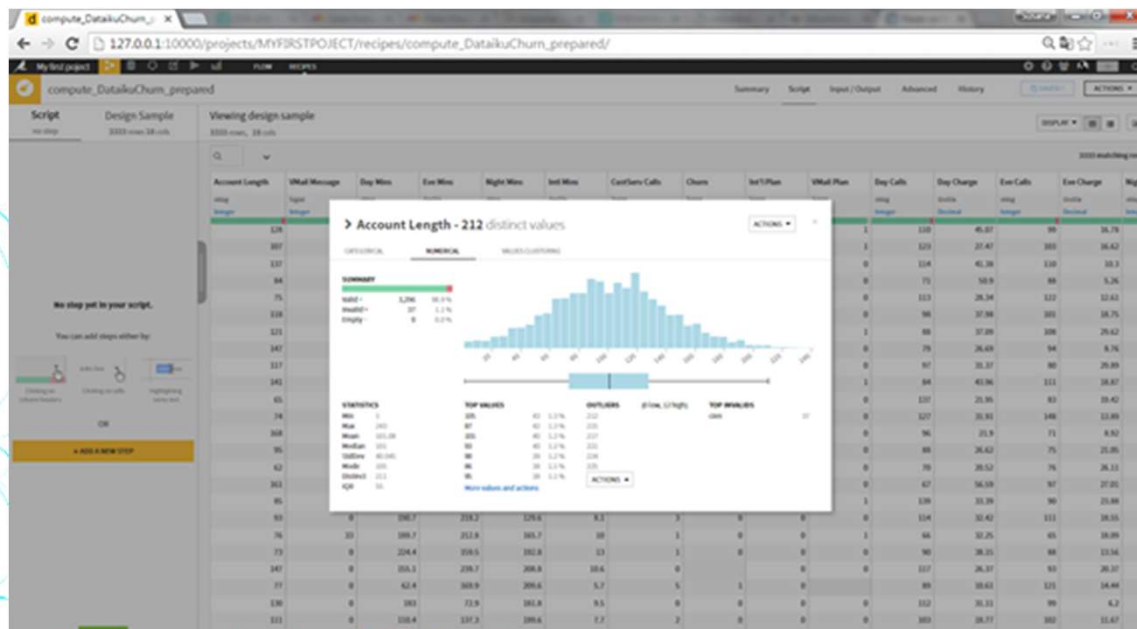
Podemos filtrar los datos, hacer sustituciones, correcciones del tipo de variable detectado, eliminación de columnas vacías y filas con valores nulos.

INTRODUCCIÓN A DATAKU DSS

Tareas Básicas con DSS

LD

1. Crear un proyecto
2. Importar un dataset
3. Vista previa y carga de los datos
4. Limpieza y preparación de los datos



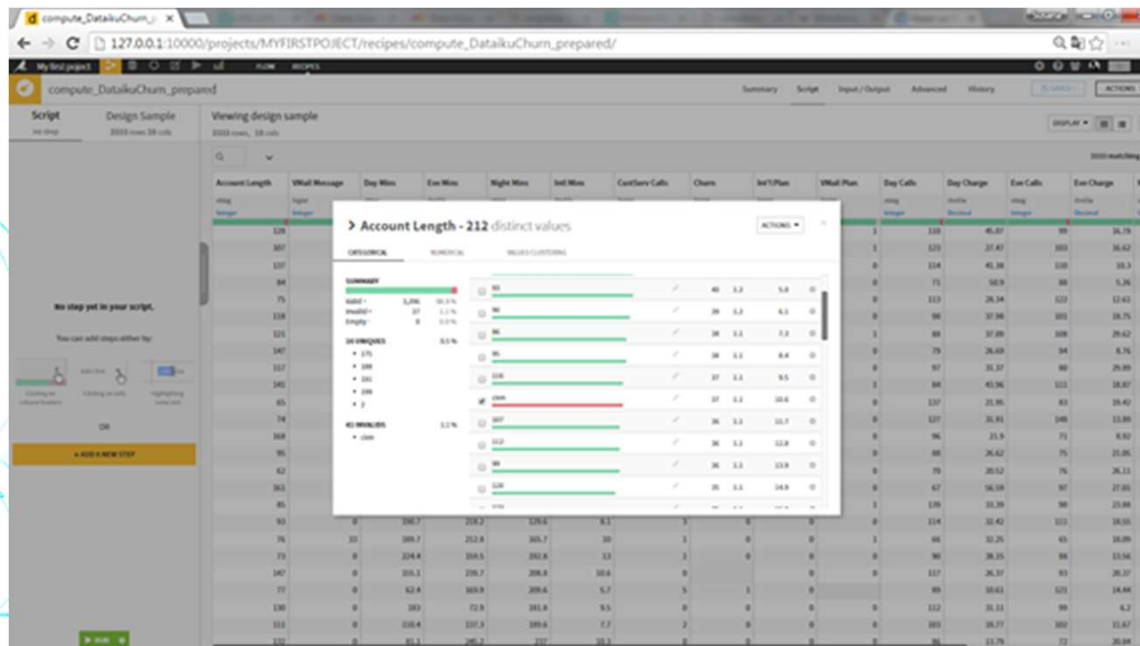
Podemos visualizar un análisis de cada una de las variables, donde se nos muestra su distribución y el análisis estadístico de la misma. Seleccionaremos la opción **Analyze** del menú de dicha columna.

INTRODUCCIÓN A DATAIKU DSS

Tareas Básicas con DSS

LD

1. Crear un proyecto
2. Importar un dataset
3. Vista previa y carga de los datos
4. Limpieza y preparación de los datos



Para realizar correcciones y modificaciones sobre los datos podemos ir añadiendo **Steps** en el proceso de limpieza.

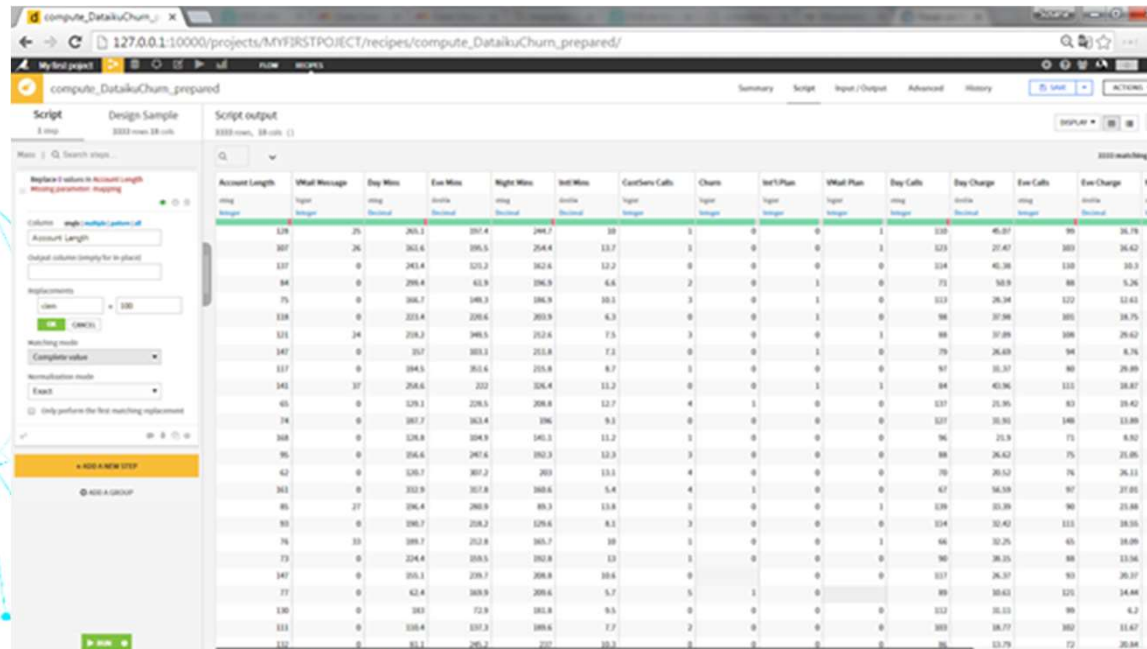
Pinchando en **Add a new step** se nos mostrarán todos los posibles nodos de acciones en la parte inferior de la pantalla.

INTRODUCCIÓN A DATAIKU DSS

Tareas Básicas con DSS

LD

1. Crear un proyecto
2. Importar un dataset
3. Vista previa y carga de los datos
4. Limpieza y preparación de los datos



Script output

Account Length	VMail Messages	Day Mins	Even Mins	Night Mins	Intl Mins	CostPerSec Calls	Churn	Intl Plan	VMail Plan	Day Calls	Day Charge	Even Calls	Even Charge	Night
328	25	245.1	107.4	244.7	38	5	0	0	1	118	45.97	98	26.78	
307	26	362.6	195.5	254.4	13.7	5	0	0	1	123	27.47	383	36.42	
117	0	243.4	121.2	162.6	12.2	0	0	0	0	114	40.38	118	30.9	
94	0	289.4	62.9	196.9	6.6	2	0	1	0	75	58.9	86	5.26	
75	0	346.7	149.3	196.9	10.1	3	0	1	0	113	26.34	122	12.61	
118	0	213.4	226.6	261.9	6.3	0	0	1	0	96	37.96	301	38.75	
121	24	218.2	148.5	212.6	7.5	3	0	0	1	88	37.89	306	26.62	
147	0	317	303.1	213.8	7.1	0	0	1	0	79	26.69	94	6.76	
117	0	184.5	301.6	215.8	8.7	1	0	0	0	97	31.37	80	26.89	
141	37	268.6	222	326.4	11.2	0	0	1	1	84	40.96	111	18.87	
45	0	129.1	228.5	208.8	12.7	4	1	0	0	137	21.96	83	19.42	
74	0	187.7	163.4	196	9.3	0	0	0	0	127	31.91	148	13.89	
368	0	128.8	104.9	140.1	11.2	1	0	0	0	96	21.9	71	8.92	
95	0	164.6	247.6	192.3	12.3	3	0	0	0	88	26.62	75	21.05	
42	0	126.7	187.2	201	13.1	4	0	0	0	70	20.52	76	26.11	
361	0	102.9	107.8	168.6	5.4	4	1	0	0	67	16.58	97	27.01	
95	27	196.4	240.9	89.3	13.8	1	0	0	1	119	31.39	90	23.88	
91	0	198.7	218.2	126.6	8.1	3	0	0	0	114	32.42	111	18.16	
76	33	189.7	212.8	165.7	10	1	0	0	1	66	32.25	45	18.09	
73	0	214.4	108.5	192.8	13	1	0	0	0	90	38.15	88	13.16	
147	0	161.1	229.7	208.8	10.6	0	0	0	0	117	26.37	93	26.17	
77	0	62.4	168.9	208.6	5.7	5	1	0		89	10.63	121	14.44	
130	0	183	72.8	181.8	9.5	0	0	0	0	112	31.11	88	6.2	
111	0	108.4	117.3	189.6	7.7	2	0	0	0	303	18.77	382	11.67	
112	0	91.1	297.2	227	10.2	0	0	0	0	36	13.79	72	20.84	

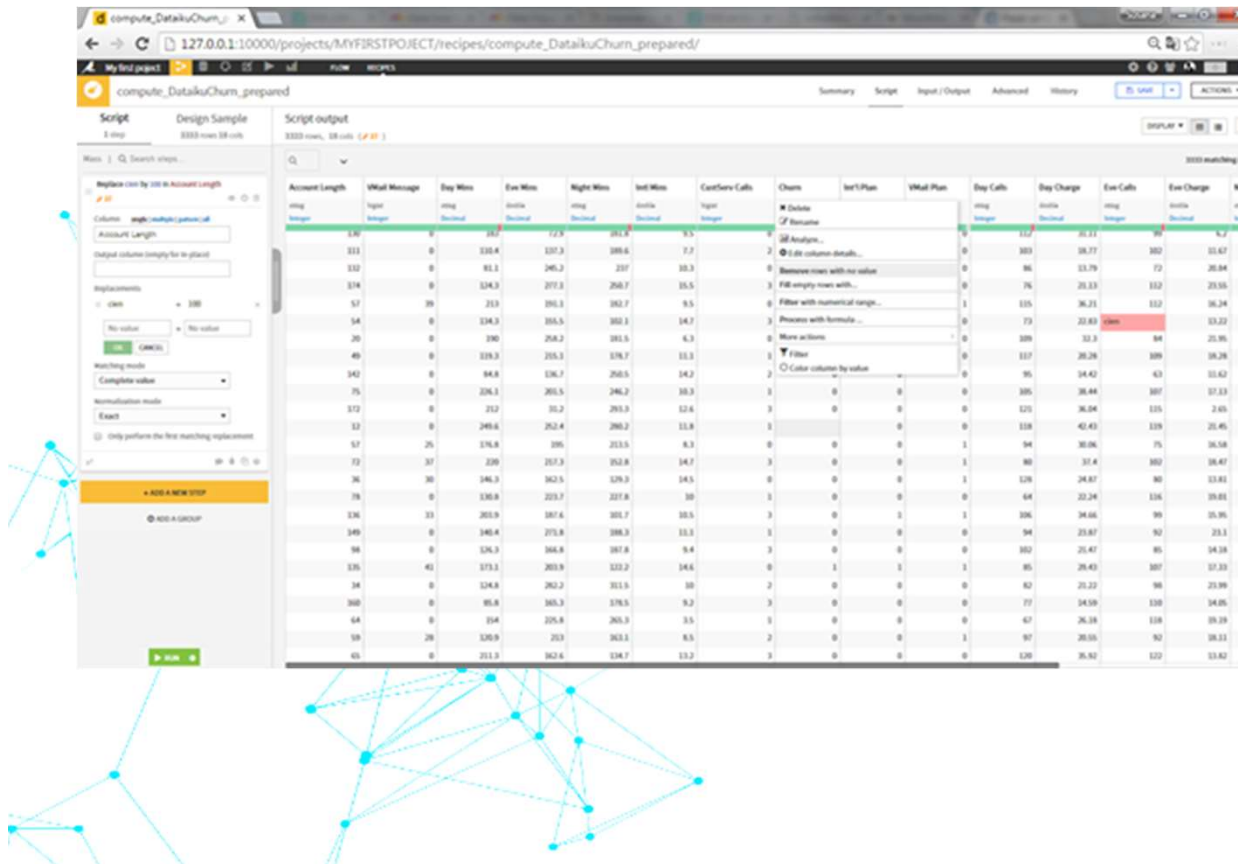
Las acciones específicas sobre una columna también se pueden seleccionar desde el menú desplegable al pinchar los nombres de las mismas.

INTRODUCCIÓN A DATAIKU DSS

Tareas Básicas con DSS

LD

1. Crear un proyecto
2. Importar un dataset
3. Vista previa y carga de los datos
4. Limpieza y preparación de los datos



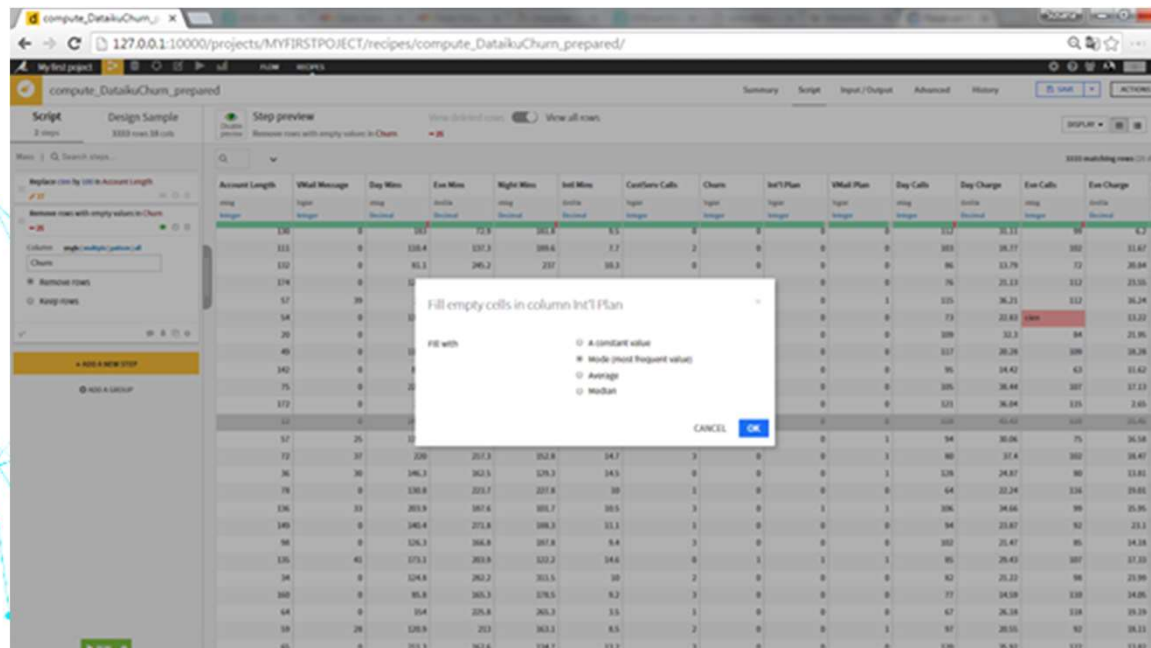
En el caso de la variable a predecir, sería conveniente eliminar aquellos registros que carezcan de valor, ya que estos datos serán utilizados para entrenar y testear el modelo. Para ello seleccionamos la opción de eliminación de aquellas filas que contengan valores nulos.

INTRODUCCIÓN A DATAIKU DSS

Tareas Básicas con DSS

LD

1. Crear un proyecto
2. Importar un dataset
3. Vista previa y carga de los datos
4. **Limpieza y preparación de los datos**



Otra opción de sustitución de valores es determinar que los valores erróneos tomen el valor del más repetido en la columna. Para ello se realizará la sustitución por la moda. Esto es conveniente para aquellas variables con un pequeño rango de valores.

INTRODUCCIÓN A DATAIKU DSS

Tareas Básicas con DSS

LD

1. Crear un proyecto
2. Importar un dataset
3. Vista previa y carga de los datos
4. Limpieza y preparación de los datos

compute_DataikuChurn_prepared

127.0.0.1:10000/projects/MYFIRSTPROJECT/recipes/compute_DataikuChurn_prepared/

My first project

compute_DataikuChurn_prepared

Script

Design Sample

3333 rows 18 cols

Step preview

View modified rows

View all rows

Summary

Script

Input/Output

Advanced

History

VIEW

ACTIONS

9 steps

3333 rows 18 cols

Match

Search steps

Replace row by 100 in Account Length

Remove row with empty values in Churn

Fill empty cells of Int'l Plan with 'V'

Fill empty cells of Int'l Plan with 'V'

Replace 2 values in Eve Calls

Column Length: Multiple patterns: all

Eve Calls

Output column (empty for to place)

Replacements

clean

clean

No value

No value

Match mode

Complete value

Normalization mode

Exact

Only perform the first matching replacement

VIEW

Script

	Night Miss	Int'l Miss	CostServ Calls	Churn	Int'l Plan	VInt'l Plan	Day Calls	Day Charge	Eve Calls	Eve Charge	Night Calls	Night Charge	Int'l Calls	Int'l Charge	
	Missed	Domestic	Integer	Integer	Integer	Integer	Missed	Domestic	Missed	Domestic	Missed	Domestic	Integer	Domestic	
1	137.3	388.6	7.7	2	0	0	0	380	18.77	380	11.67	385	8.53	4	2.08
2	246.2	237	18.3	0	0	0	0	96	13.79	72	26.84	125	38.67	2	2.76
3	271.1	268.7	15.5	3	0	0	0	76	21.13	112	23.55	125	11.28	5	4.19
4	391.1	382.7	8.5	0	0	0	1	135	36.21	112	36.24	125	8.22	3	2.57
5	105.5	382.1	14.7	3	0	0	0	73	21.83	108	12.22	48	4.19	4	1.97
6	268.2	381.5	4.3	0	0	0	0	309	32.3	84	21.95	380	8.17	6	1.7
7	271.1	176.7	11.1	1	0	0	0	117	26.26	109	18.26	90	8.04	1	3
8	126.7	260.5	14.2	2	0	0	0	95	14.42	43	11.62	148	11.27	6	3.83
9	281.5	246.2	18.3	1	0	0	0	105	36.44	107	17.13	98	11.08	5	2.76
10	31.2	281.3	12.6	3	0	0	0	121	36.84	125	2.65	78	13.2	10	3.4
11	395	211.5	8.3	0	0	0	1	94	38.06	75	36.58	124	9.61	4	2.24
12	257.3	352.8	14.7	3	0	0	1	80	37.4	380	18.47	71	6.88	6	1.97
13	162.5	126.3	14.5	0	0	0	1	128	24.87	80	13.81	109	5.62	6	1.92
14	223.7	227.8	18	1	0	0	0	64	12.24	124	19.81	108	18.25	5	2.7
15	187.6	181.7	18.5	3	0	1	1	106	34.66	99	16.95	107	4.58	6	2.84
16	271.8	188.3	11.1	1	0	0	0	94	21.87	92	23.1	108	8.47	9	3
17	168.8	167.8	3.4	3	0	1	0	102	11.47	91	14.18	125	6.46	2	2.14
18	281.9	112.2	14.6	0	1	1	1	85	28.42	107	17.13	79	5.5	15	3.94
19	282.2	111.5	14	2	0	0	0	82	21.22	98	25.96	78	14.12	4	2.7
20	360.3	178.5	8.2	3	0	0	0	77	14.18	118	14.05	82	8.03	4	2.48
21	125.8	261.3	3.5	1	0	0	0	47	26.18	118	19.19	86	11.94	3	0.95
22	213	363.1	8.5	2	0	0	1	87	28.05	82	18.11	124	7.34	5	2.3
23	162.6	124.7	13.2	3	0	0	0	120	16.92	122	13.82	118	6.86	5	1.94
24	124.6	242.2	7.4	2	0	0	0	110	11.79	74	11.44	127	18.19	5	2
25	171.3	141.7	8.8	4	1	0	0	118	11.05	111	18.05	81	6.48	3	0.78

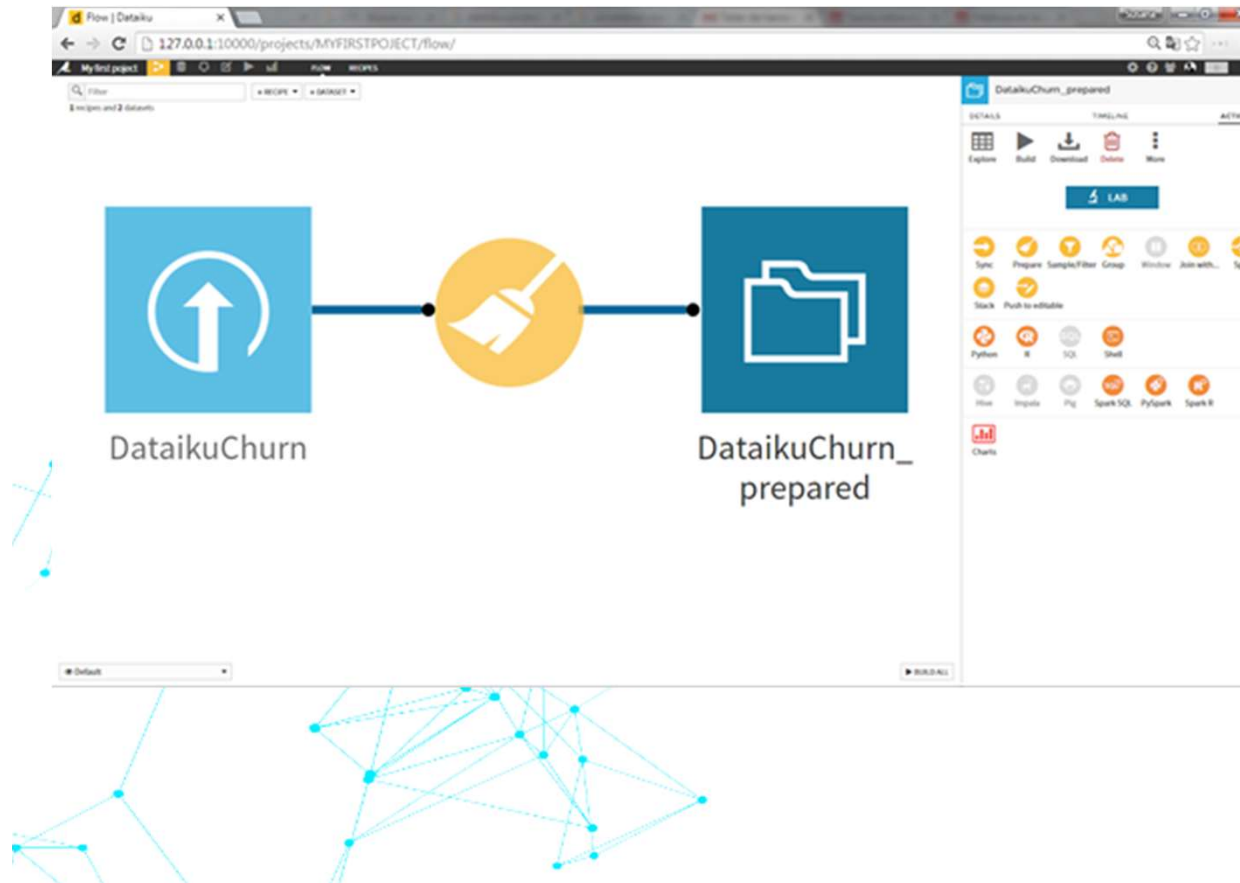
Los cambios se pueden realizar sobre varias columnas simultáneamente, seleccionando **Multiple** en vez de **Single**, y se pueden añadir tantas columnas como quieras, sobre las que se aplicarán los cambios.

INTRODUCCIÓN A DATAIKU DSS

Tareas Básicas con DSS

LD

1. Crear un proyecto
2. Importar un dataset
3. Vista previa y carga de los datos
4. Limpieza y preparación de los datos



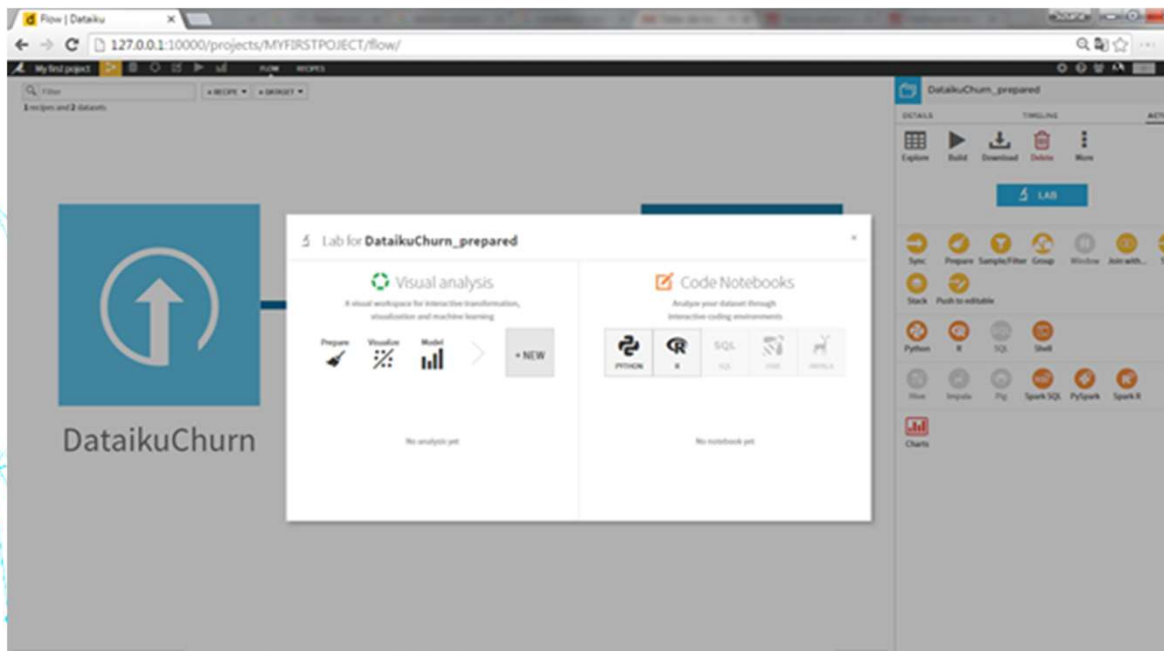
Una vez se han determinado todos los cambios que se quieren llevar a cabo, se ejecutan. Para ello basta con hacer **Run**. Así se nos crea el nuevo **dataset** con todos los cambios.

INTRODUCCIÓN A DATAIKU DSS

Tareas Básicas con DSS

MP

1. Crear nuevo modelo
2. Definir modelos y parámetros
3. Evaluación y comparación de modelos



Podemos crear un modelo predictivo sobre el nuevo **dataset**. Para ello haremos clic sobre el nuevo **dataset** y seleccionaremos la opción de **LAB**, que se nos muestra como opción.

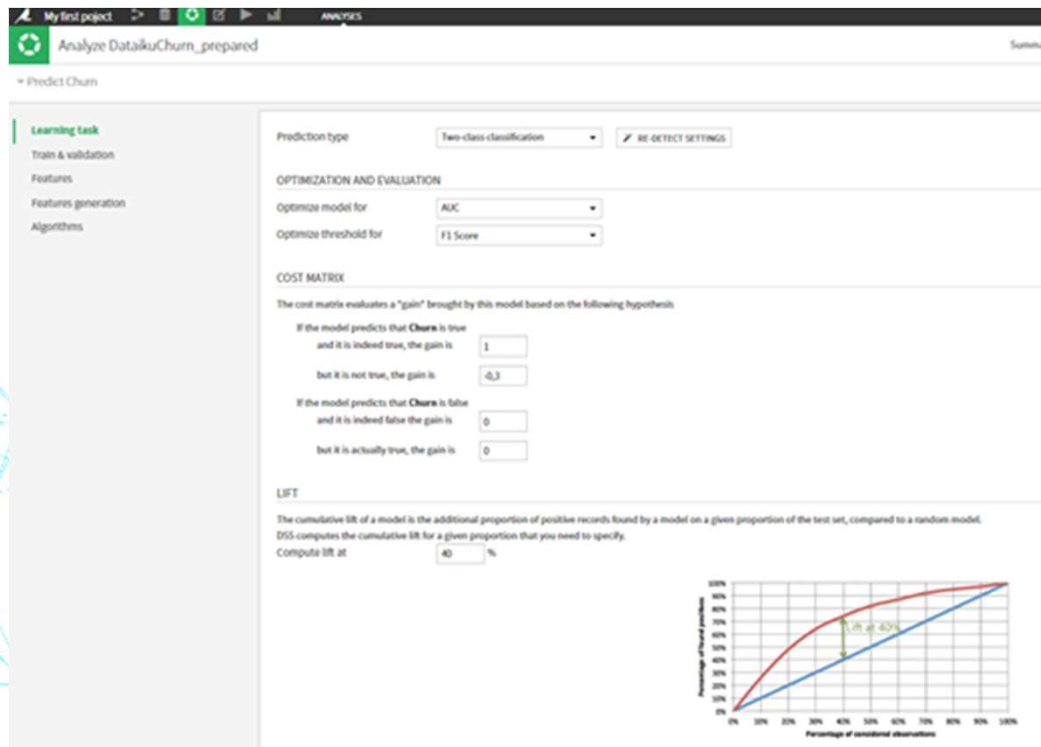
Creamos un nuevo modelo, en caso de existir alguno nos aparecería listado debajo del cuadro de **Visual Analysis**.

INTRODUCCIÓN A DATAIKU DSS

Tareas Básicas con DSS

MP

1. Crear nuevo modelo
2. **Definir modelos y parámetros**
3. Evaluación y comparación de modelos



My first project

Analyze DataikuChurn_prepared

Predict Churn

Learning task

Train & validation

Features

Features generation

Algorithms

Prediction type: Two-class classification

RE-DETECT SETTINGS

OPTIMIZATION AND EVALUATION

Optimize model for: AUC

Optimize threshold for: F1 Score

COST MATRIX

The cost matrix evaluates a "gain" brought by this model based on the following hypothesis:

If the model predicts that **Churn** is true and it is indeed true, the gain is: 1

but it is not true, the gain is: -0.3

If the model predicts that **Churn** is false and it is indeed false the gain is: 0

but it is actually true, the gain is: 0

LIFT

The cumulative lift of a model is the additional proportion of positive records found by a model on a given proportion of the test set, compared to a random model. DSS computes the cumulative lift for a given proportion that you need to specify.

Compute lift at: 40 %

Percentage of total positives

Percentage of selected observations

Lift at 40%

Se pueden seleccionar las métricas, hacer filtros sobre los datos a analizar.

Antes de entrenar el modelo, en **Settings** se puede decidir qué modelos entrenar y modificar los parámetros de los mismos.

INTRODUCCIÓN A DATAIKU DSS

Tareas Básicas con DSS

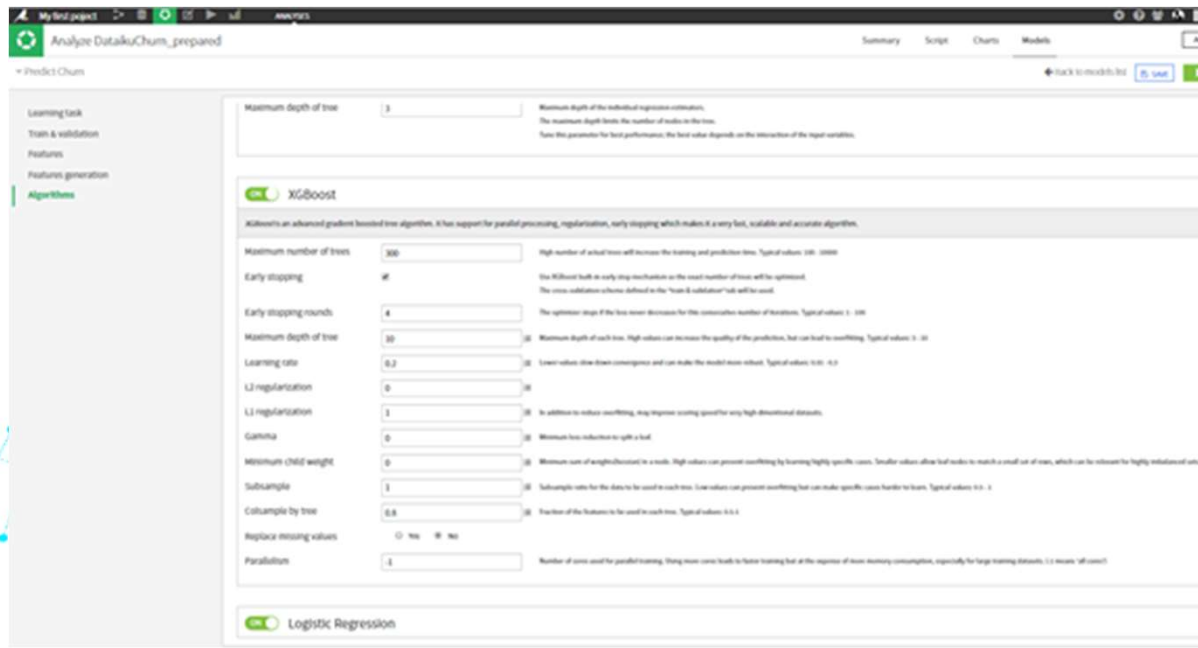
MP

1. Crear nuevo modelo
2. **Definir modelos y parámetros**
3. Evaluación y comparación de modelos

Hay varios algoritmos disponibles para la predicción:

- Random forest
- Gradient boosted tree
- Xgboost
- Logistic regresión
- Decision tree
- Support vector machine
- Stochastic gradient descent

También se puede añadir un modelo propio en Python

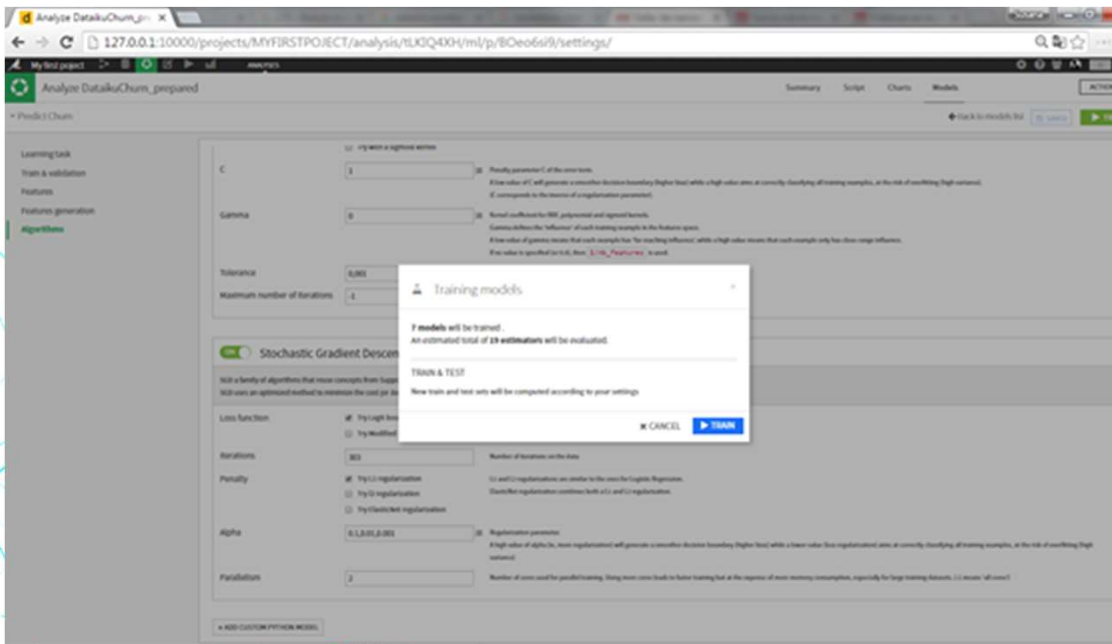


INTRODUCCIÓN A DATAIKU DSS

Tareas Básicas con DSS

MP

1. Crear nuevo modelo
2. **Definir modelos y parámetros**
3. Evaluación y comparación de modelos



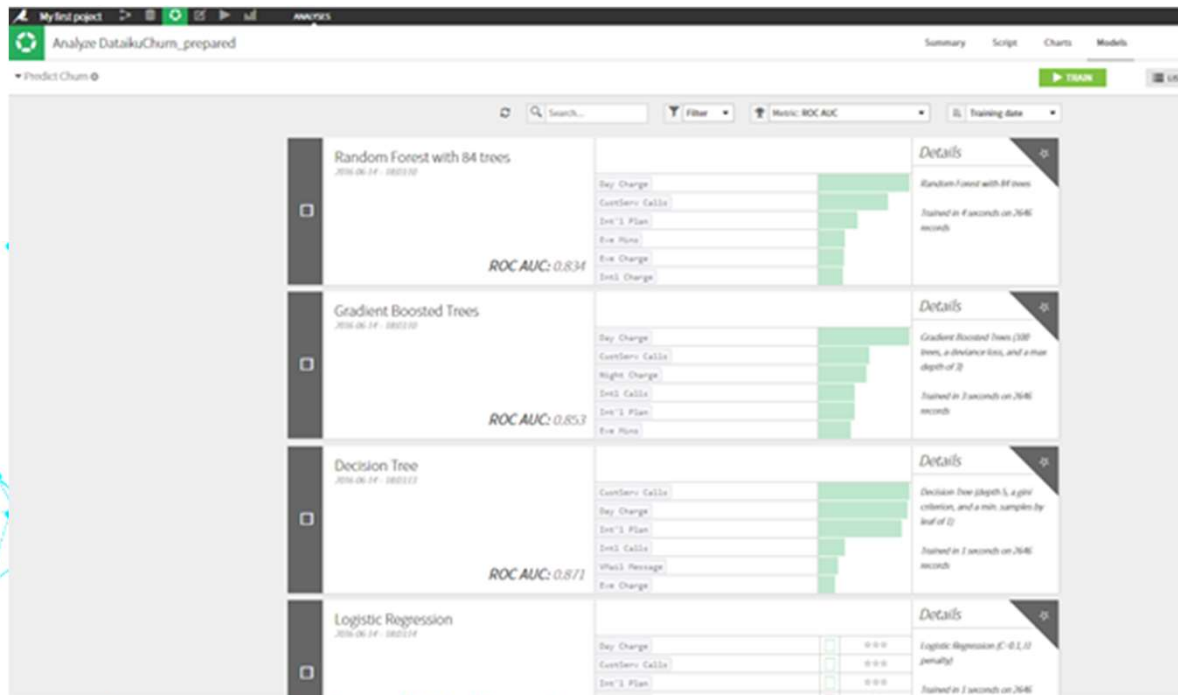
Cuando hemos seleccionado los algoritmos a utilizar podemos proceder a entrenarlos, simplemente pulsando ***Train***.

INTRODUCCIÓN A DATAIKU DSS

Tareas Básicas con DSS

MP

1. Crear nuevo modelo
2. Definir modelos y parámetros
3. Evaluación y comparación de modelos



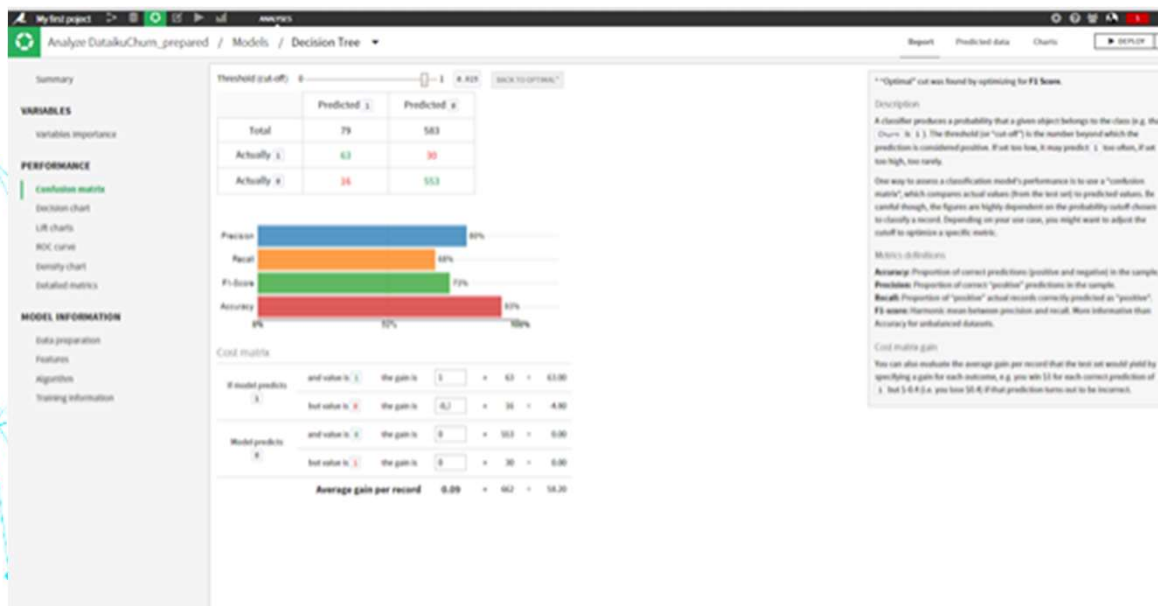
Una vez entrenados los modelos, podemos compararlos basados en su performance, y así determinar cuál nos resulta más útil para nuestro análisis de predicción.

INTRODUCCIÓN A DATAIKU DSS

Tareas Básicas con DSS

MP

1. Crear nuevo modelo
2. Definir modelos y parámetros
3. Evaluación y comparación de modelos



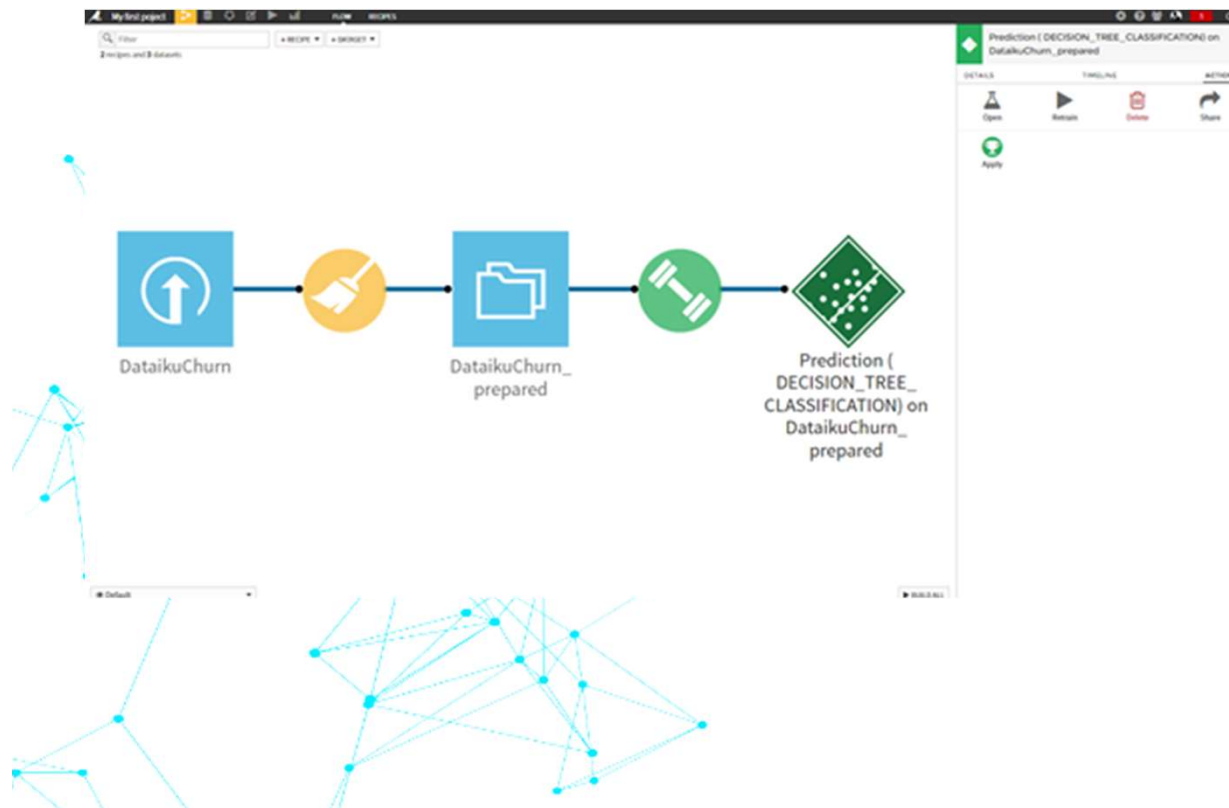
Para cada modelo disponemos de la información referente a las variables más importantes, el rendimiento, etc.

INTRODUCCIÓN A DATAIKU DSS

Tareas Básicas con DSS

MP

1. Crear nuevo modelo
2. Definir modelos y parámetros
3. Evaluación y comparación de modelos



Una vez seleccionado el mejor modelo, para poder utilizarlo hacemos **deploy**.

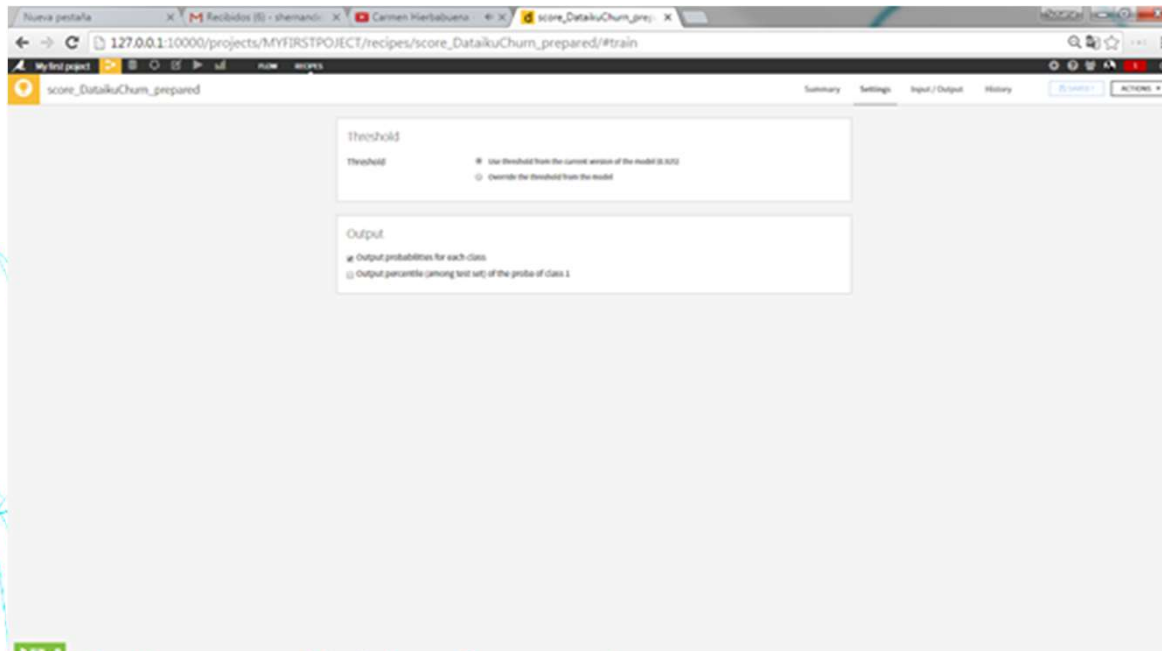
Desde el flujo de trabajo del modelo podemos seleccionar el modelo y aplicarlo sobre un **dataset** sobre el que queremos realizar la predicción. Para ello seleccionamos **Apply**.

INTRODUCCIÓN A DATAIKU DSS

Tareas Básicas con DSS

MP

1. Crear nuevo modelo
2. Definir modelos y parámetros
3. Evaluación y comparación de modelos



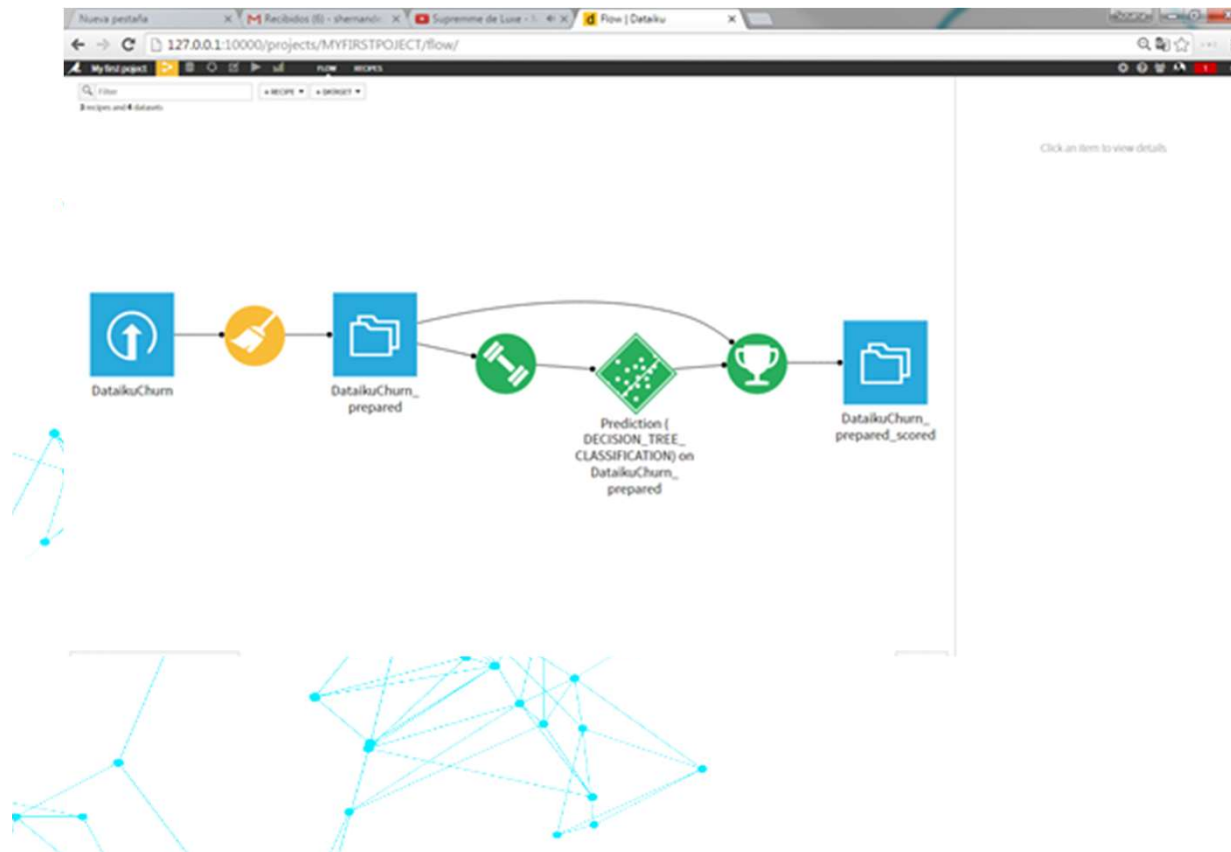
Al seleccionar **output probability for each class** nos permitirá visualizar la distribución de las predicciones. Una vez seleccionado hacemos clic en **RUN**.

INTRODUCCIÓN A DATAIKU DSS

Tareas Básicas con DSS

MP

1. Crear nuevo modelo
2. Definir modelos y parámetros
3. **Evaluación y comparación de modelos**



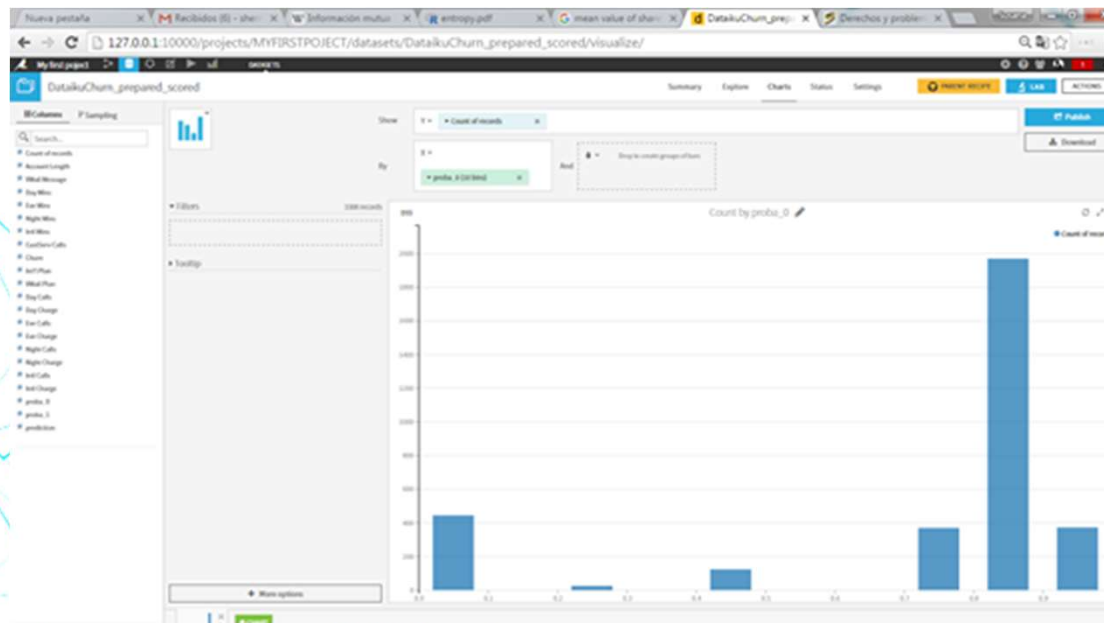
Se nos generará un nuevo **dataset** con 3 nuevas columnas. En este caso la de las probabilidades de las posibles predicciones y la columna con la predicción.
(Prediction,proba_0, proba_1)

INTRODUCCIÓN A DATAIKU DSS

Tareas Básicas con DSS

V

1. Crear visualización



Para visualización de resultados, hacer click en **Charts**.

Por ejemplo, crear un histograma con la distribución de probabilidad, o comparar valores reales con valores predichos.

Conecta Empleo

