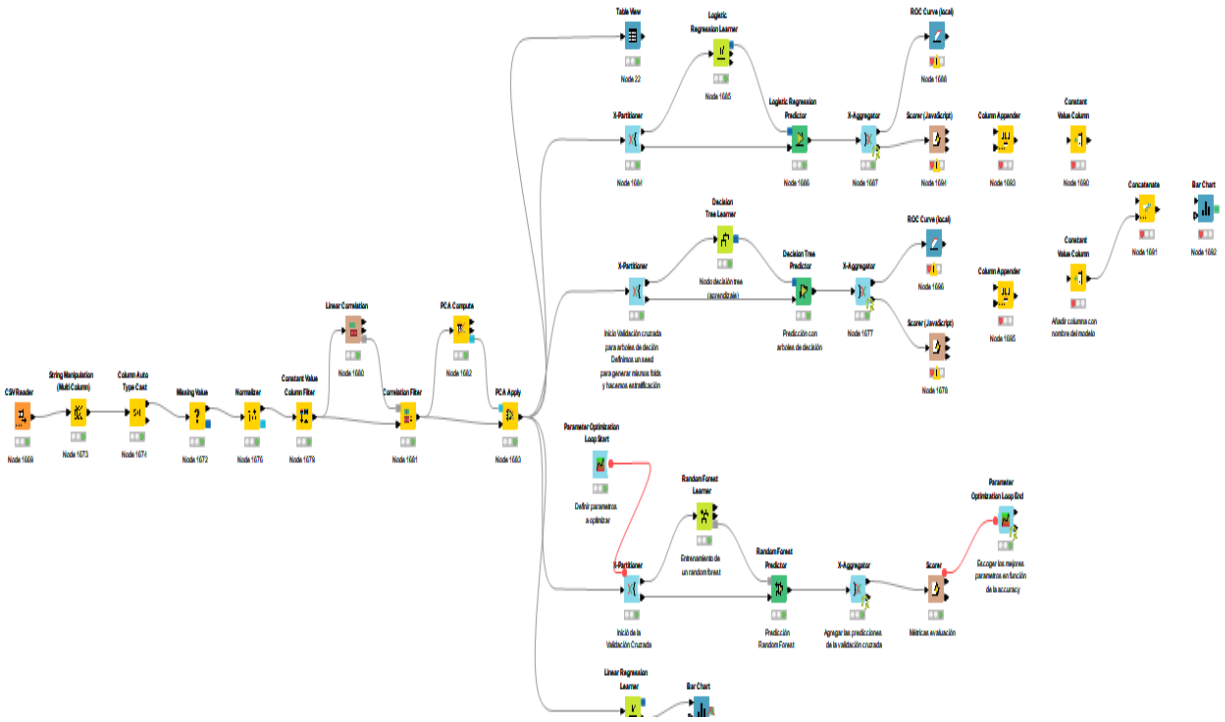
Matias Dávila Winder
Ampliación de la estadística
Para Data Science

UNIVERSITAT Carlemany

Evaluación y comparación de Modelos básicos



Profesor:

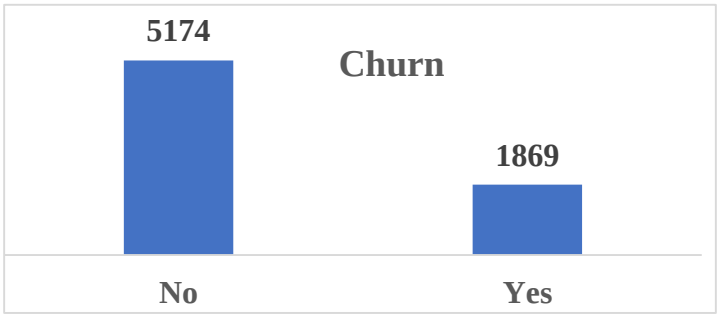
Francesc Busquet

Matias Dávila Winder
Ampliación de la estadística
Para Data Science

UNIVERSITAT Carlemany

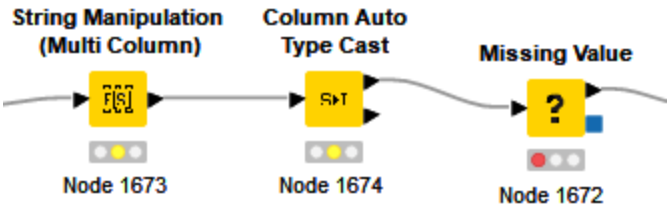
Introducción

Durante este trabajo se estudió y analizo una base de datos, en la cual debíamos ordenar, limpiar y analizar. En esta base de datos podemos ver que trata de una empresa de telecomunicaciones, en la cual encontramos los servicios que ofrece y los contratos con sus respectivos clientes, en los cuales se observa que tipos de servicios contrata cada persona, línea de teléfono, internet, servicios de streaming, entre otros. Luego nos encontramos con la duración de este contrato, el cargo que tiene mensualmente, el cargo total y la tasa de cancelación(Churn), para el caso de estos valores se ha buscado si existen clientes con celdas vacías en estas columnas y si es el caso se remplazaron por un 0 o "null".



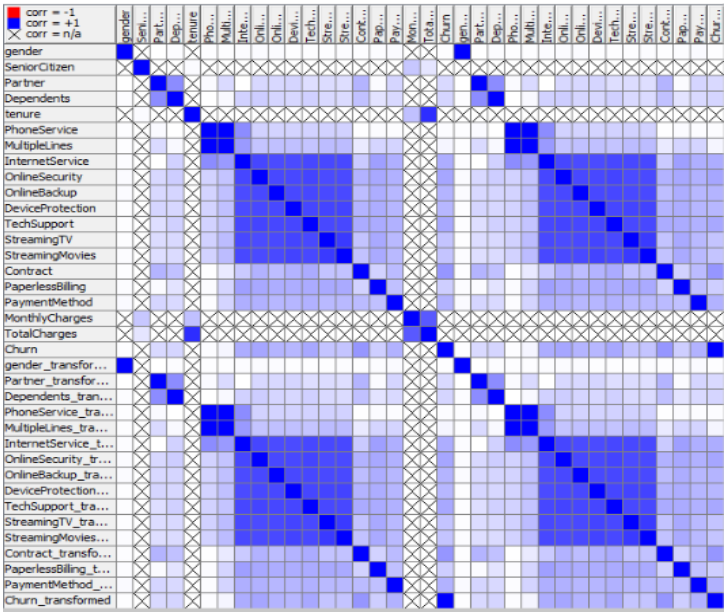
En esta actividad se nos pide predecir la variable Churn, como información básica se conoce que, un **27%** de los datos corresponden a "Yes" y **73%** corresponden a "No". En este caso para llevarlo a un modelo y poder trabajarlo de mejor manera, en general se realiza el cambio a binario donde **Yes = 1 y No = 0**. Para estos casos lo más común es utilizar regresiones lineales o logísticas.

Evaluamos el proceso a través de 3 modelos distintos, una regresión logística, decitional tree y a través de optimización. Para esto se debió hacer un proceso previo, el cual entendemos como el procesamiento de datos en donde identificamos los missing values, cambiándolos de valor para



que el computador los pueda entender.

Luego se normalizan los datos y revisamos la correlacion que existe entre los datos debido a que si queremos que los modelos planteados funcionen de manera correcta los datos no pueden estar correlacionados.



Matias Dávila Winder
Ampliación de la estadística
Para Data Science

Una vez finalizado el procesamiento de los datos debemos ajustar los modelos para que funcionen correctamente. A continuación, veremos los resultados obtenidos en cada modelo con los cuales seremos capaces de generar una conclusión y poder elegir cual es el que mejor funciona para esta actividad.

- Logistic Regression

Chrun
Confusion Matrix

	No (Predicted)	Yes (Predicted)	
No (Actual)	4642	532	89.72%
Yes (Actual)	871	998	53.40%
	84.20%	65.23%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
80.08%	19.92%	0.458	5640	1403

- Decision Tree

Chrun
Confusion Matrix

	No (Predicted)	Yes (Predicted)	
No (Actual)	4306	866	83.26%
Yes (Actual)	902	964	51.66%
	82.68%	52.68%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
74.88%	25.12%	0.351	5270	1768

- Optimización

Churn \ Pr...	No	Yes
No	4690	484
Yes	960	909

Correct classified: 5.599 Wrong classified: 1.444

Accuracy: 79,497% Error: 20,503%

Cohen's kappa (κ): 0,428%

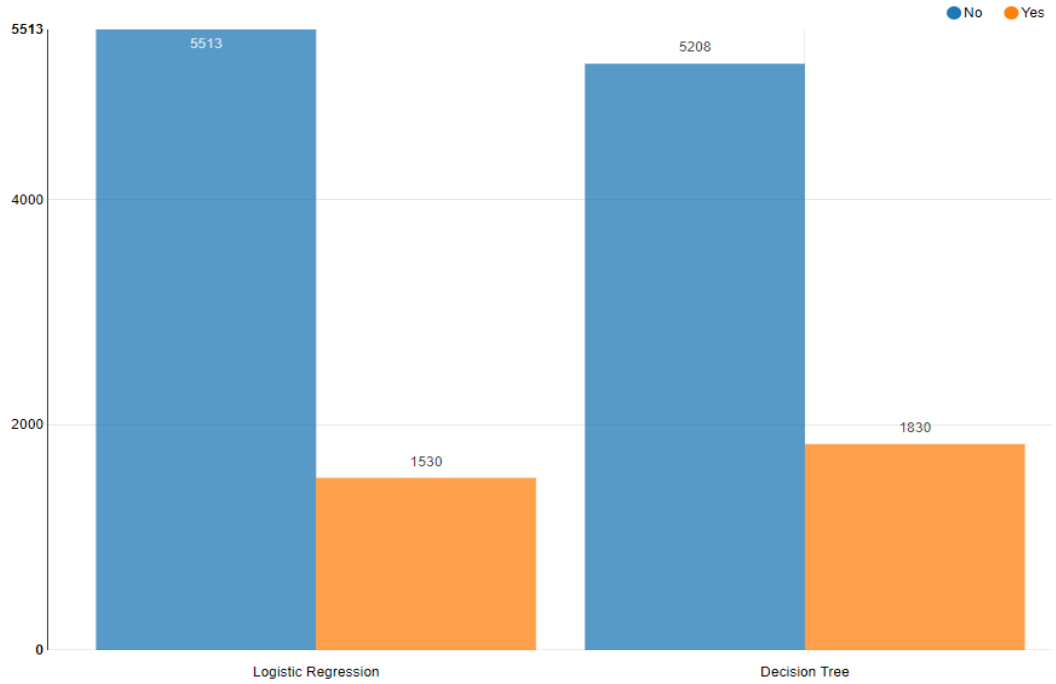
Matias Dávila Winder
Ampliación de la estadística
Para Data Science

UNIVERSITAT Carlemany

Conclusión

En conclusión, durante esta actividad hemos realizado los pasos a seguir para estudiar bases de datos, se separó cada variable en una respectiva columna, se eliminó los separadores (,) y se hace el intercambio entre “Missing value” a un valor Null o 0.

Al estudiar los resultados obtenidos en cada modelo, podemos ver que para este caso la predicción de la variable chrun, resulto mejor en la regresión logística. Con este modelo logramos una precisión de **84,2%** en la predicción de la variable mientras que solo posee un error general de **19,92%**. Si bien los otros modelos logran un resultado bastante similar es mejor quedarse con el que es más acertado siempre.



En lo particular esta actividad fue bastante desafiante para mí, conocía poco la plataforma KNIME, es sencilla de ocupar la verdad, esta entrega mucha información la cual en un inicio el me costó procesar, ya que debia entender, explorar y familiarizarme con ella.

Buena actividad, ya te deja con la posibilidad de explorar el mundo de machine learning que es esencial para la ciencia de datos.

.