

ID: 99905

# MBD\_Módulo 5: Caso Práctico Individual

---

## Churn Prediction

Kaggle: WA\_Fn-UseC\_-Telco-Customer-Churn.csv



# Variables DataSet

- **CustomerID.** Customer ID
- **gender.** Si el cliente es Male o Female
- **SeniorCitizen.** Si el cliente es “senior citizen” o no (1, 0)
- **Partner.** Si el cliente tiene un distribuidor o no (Yes, No)
- **Dependents.** Si el cliente tiene “dependents” (Yes, No)
- **tenure.** Número de meses en la empresa.
- **Contract.** Tipo de contrato.
- **PaperlessBilling.** Factura electrónica (Yes, No).
- **PaymentMethod.** Método de pago
- **MonthlyCharges.** Cuota mensual.
- **TotalCharges.** Total cantidad facturada.
- **Churn.** Variable objetivo. (Yes or No)

- **PhoneService.** Si tiene servicio telefónico o no (Yes, No)
- **MultipleLines.** Si tiene múltiples líneas o no (Yes, No, No phone service)

- **InternetService.** Servicio de datos (DSL, Fiber optic, No)
- **OnlineSecurity.** Servicio añadido
- **OnlineBackup.** Servicio añadido
- **DeviceProtection.** Servicio añadido
- **TechSupport.** Servicio añadido
- **StreamingTV.** Servicio añadido
- **StreamingMovies.** Servicio añadido

Todos tienen valores Yes, No, No internet service

# Variables Extras

**DiffCharges** para estudiar la diferencia entre **TotalCharges** y **tenure-MonthlyCharges**. Se interpreta como un cambio de tarifa.

Interesante categorizar aquellos que han cambiado a mejores servicios (up selling/cross selling) o todo lo contrario.

**DiffCharges** =  $\text{TotalCharges} / \text{tenure-MonthlyCharges}$ .

**TheoMonthlyCharges** a partir del estudio realizado de precios individuales de cada servicio contratado:

- Servicio Voz => 20.
- Servicio Voz añadido multi línea => +5 (es decir, voz + multi línea = 25)
- Servicio DSL => 25 (Voz + DSL = 45)
- Por cada servicio añadidos OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport => +5
- Por cada servicio añadidos StreamingTV, StreamingMovies => +10
- En el Servicio de Fibra, va incluida la Voz => 70

Cliente **Fibra** con todo contratado = 110

Esta variable nos permite visualizar de un forma muy fácil los servicios a partir de las cuotas.

Además, quita el ruido de decimales de cuotas mensuales.

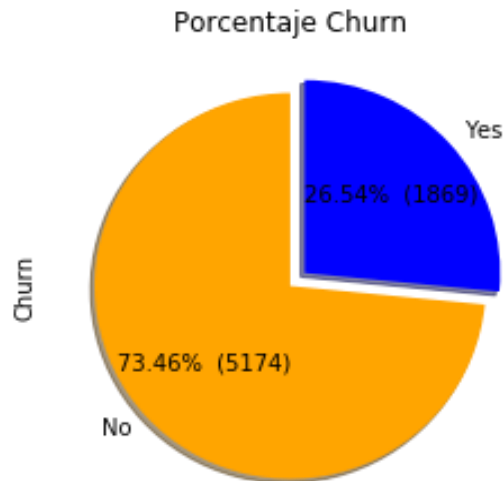
**ICEMD**

INSTITUTO ECONOMÍA  
DIGITAL



**ESIC**

# Análisis Exploratorio



Dataset desbalanceado.

¡¡Obliga a ejecutar técnicas para paliarlo!!

SMOTE

Near Miss

Cross Validation

Para Optimización de modelo, junto con Cross Validation (CV)

Grid SearchCV

Random SearchCV

RFECV

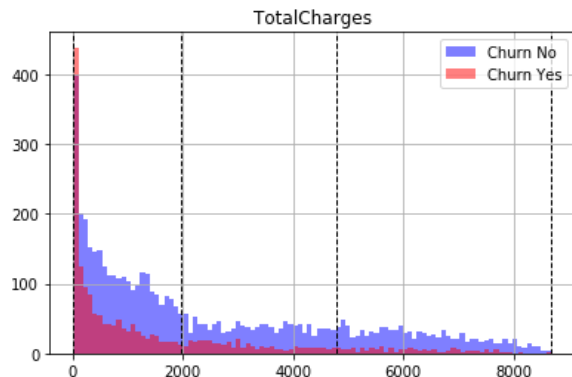
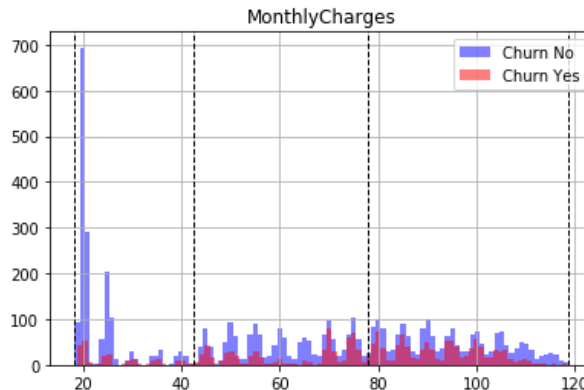
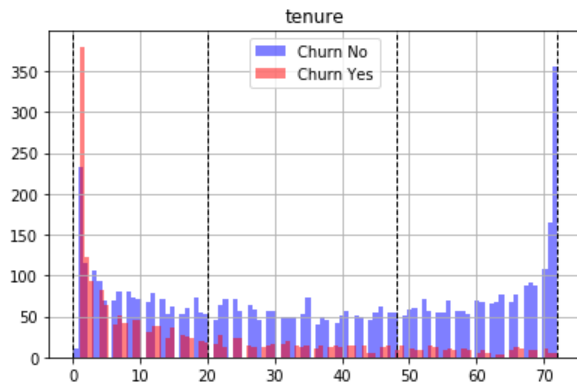
# ICEMD

INSTITUTO ECONOMÍA  
DIGITAL



ESIC

# Análisis Exploratorio. Variables Numéricas



## Intervalos Discretos

Tenure:

[0.0, 20.0, 48.0, 72.0]

MonthlyCharges:

[18.25, 42.4, 77.8, 118.75]

TotalCharges:

[18.8, 1980.3, 4786.15, 8684.8]

Crearemos variables discretizadas:

- tenure\_bin [1 2 3]
- MonthlyCharges\_bin [1 2 3]
- TotalCharges\_bin [1 2 3]

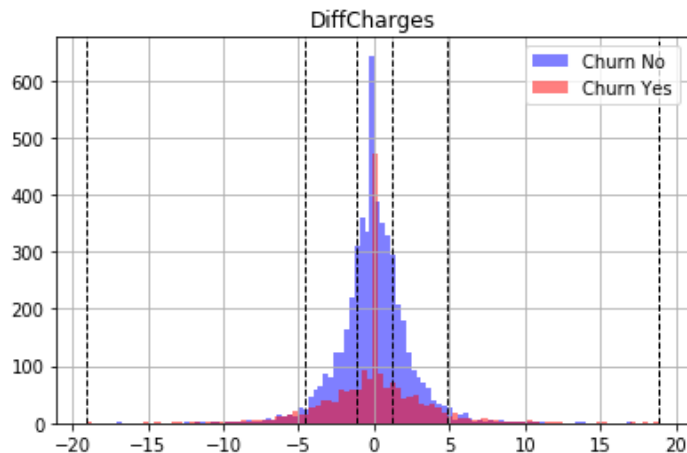
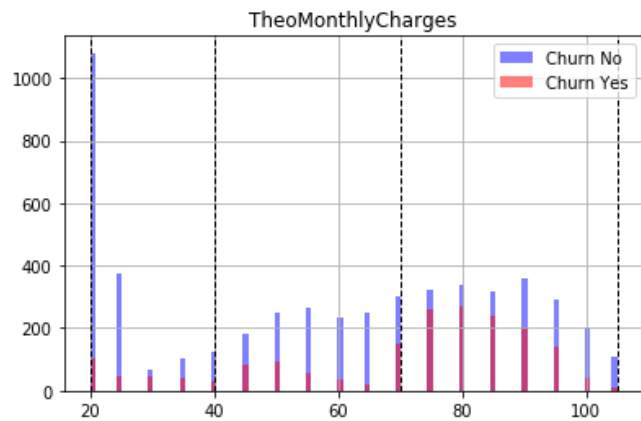
# ICEMD

INSTITUTO ECONOMÍA  
DIGITAL



ESIC

# Análisis Exploratorio. Variables Numéricas



## Intervalos Discretos

TheoMonthlyCharges:

[20.0, 40.0, 70.0, 105.0]

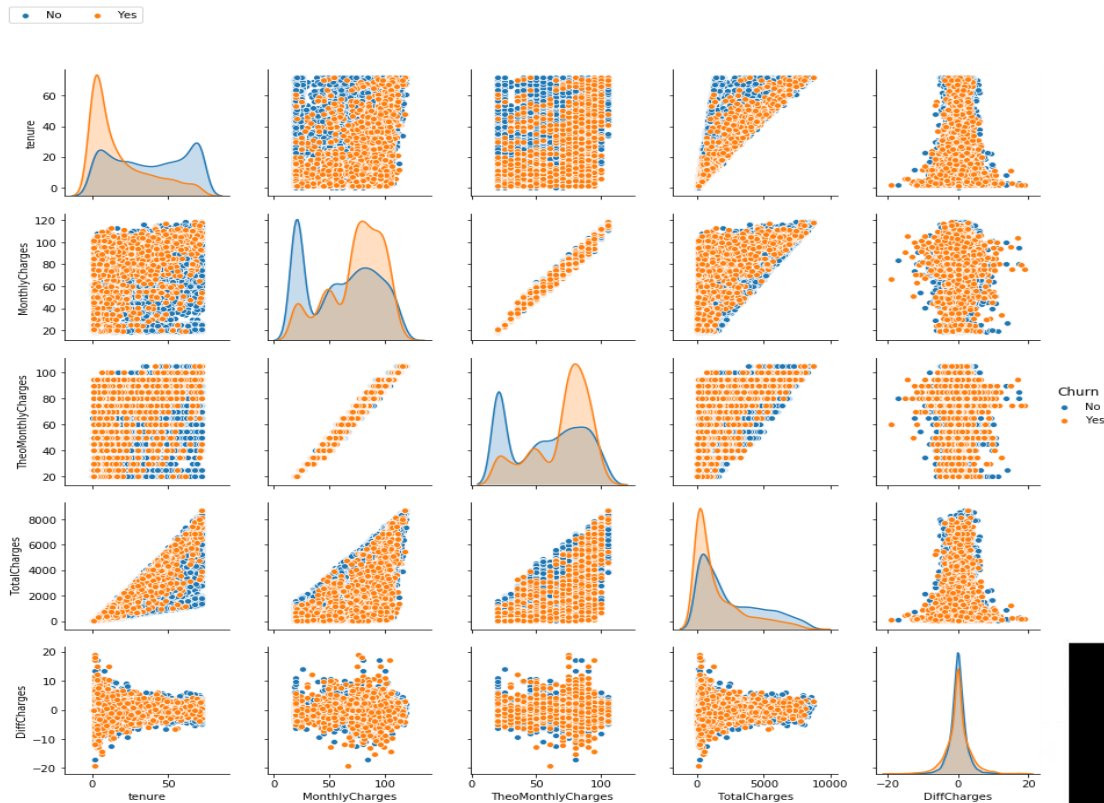
DiffCharges:

[-19.125, -4.5643, -1.2257, 1.2062, 4.82, 18.9]

Crearemos variables discretizadas:

- TheoMonthlyCharges\_bin [1 2 3]
- DiffCharges\_bin [1 2 3 4 5]

# Análisis Exploratorio. Variables Numéricas



Cuotas inferior de los 25 (**Voz**) son bastante fieles.

Clientes con cuotas altas (**Fibra**) mayor tasa de abandono.

Clientes nuevos (**tenure <3**) mayor tasa de abandono

**DiffCharges** presenta una distribución muy normal

Mucha concentración en clientes que **no cambian de contrato**, siendo la variación al alza y a la baja de 5, es decir, de servicios añadidos de Voz (**MultipleLines**) o datos (**Online**, **TechSupport**, etc..).

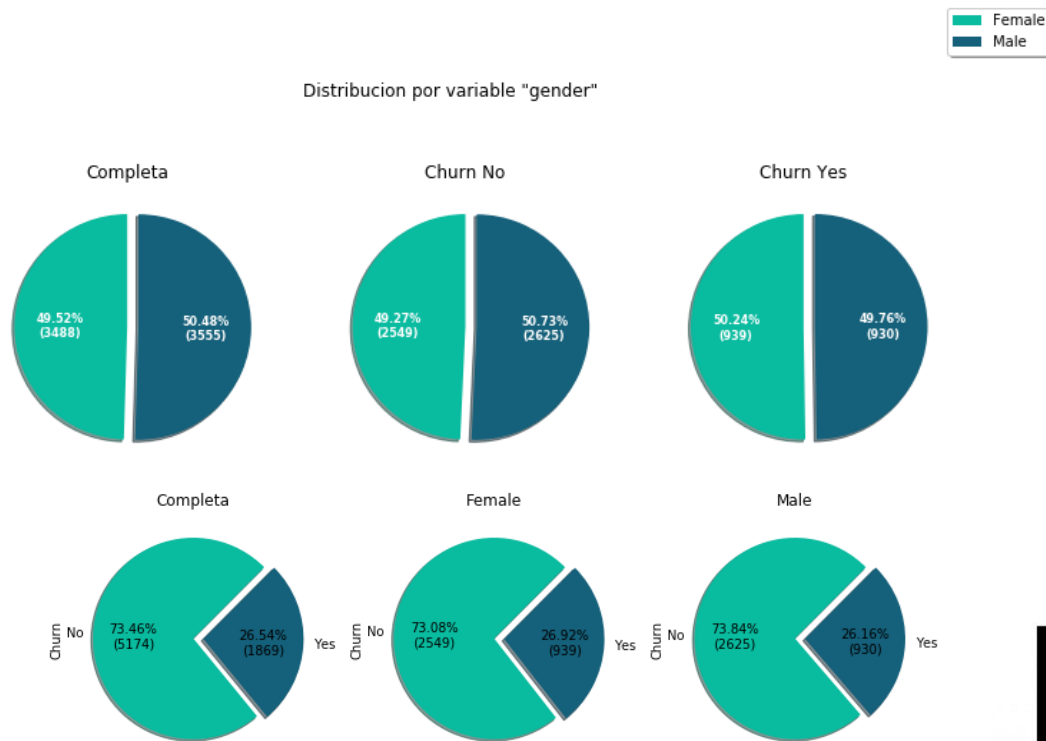
# ICEMD

INSTITUTO ECONOMÍA  
DIGITAL



ESIC

# Análisis Exploratorio. Variables categóricas



**gender.** Prácticamente, existe el mismo número de clientes de sexo masculino y femenino. En concreto, 67 registros más del sexo masculino. Esta relación se mantiene tanto en aquellos que abandonan como los que no.

Fijarse, además, que la relación de Churn por cada género es muy parecido al global.

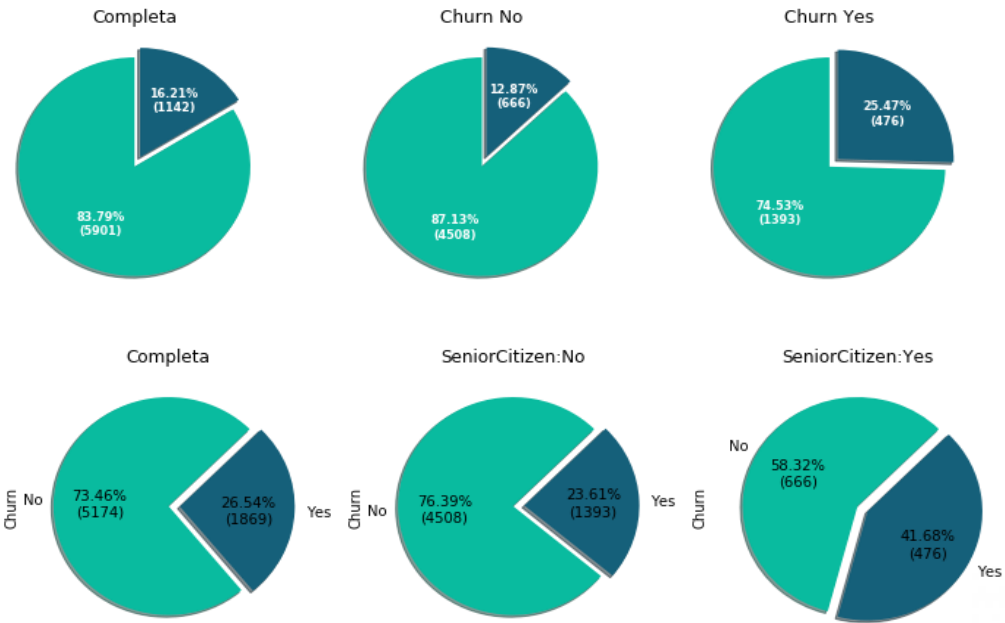
Por todo ello, parece que ésta variable no es significativa como para tenerla en cuenta en la predicción.



# Análisis Exploratorio. Variables categóricas

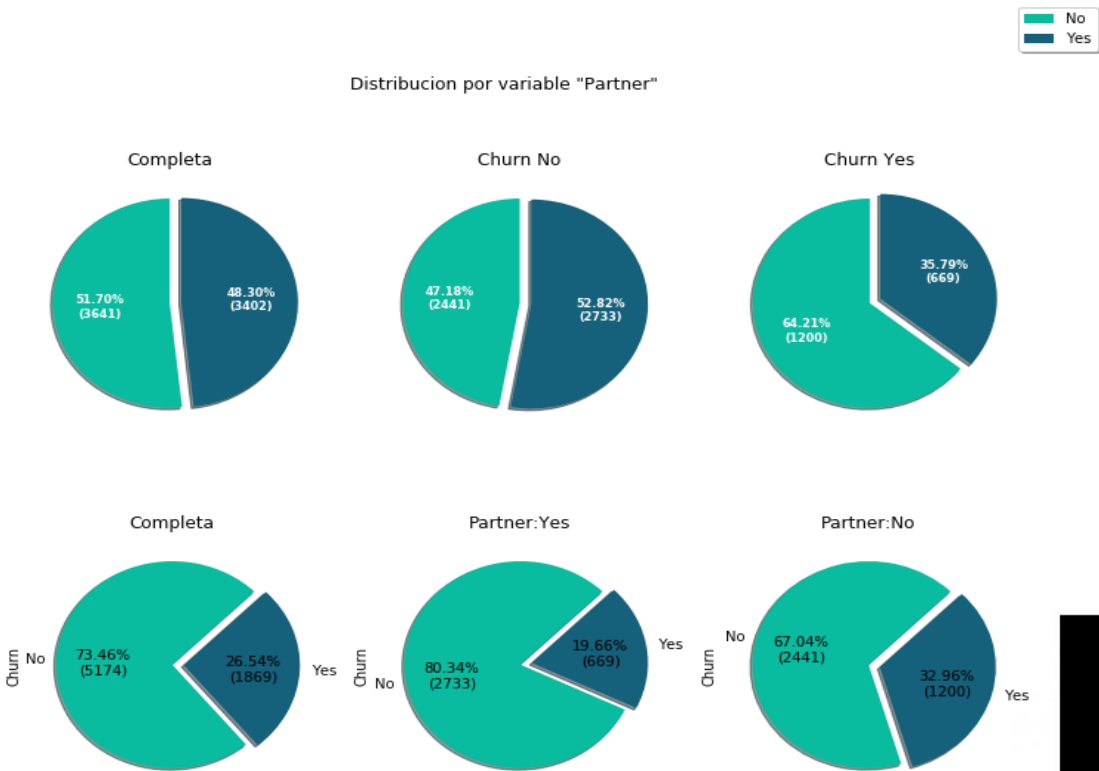


Distribucion por variable "SeniorCitizen"



**SeniorCitizen.** Los “senior citizen” tienen bastante menos representación teniendo una tasa de abandono superior al global. Sí parece que ser Senior Citizen o no puede condicionar la tasa de abandono.

# Análisis Exploratorio. Variables categóricas

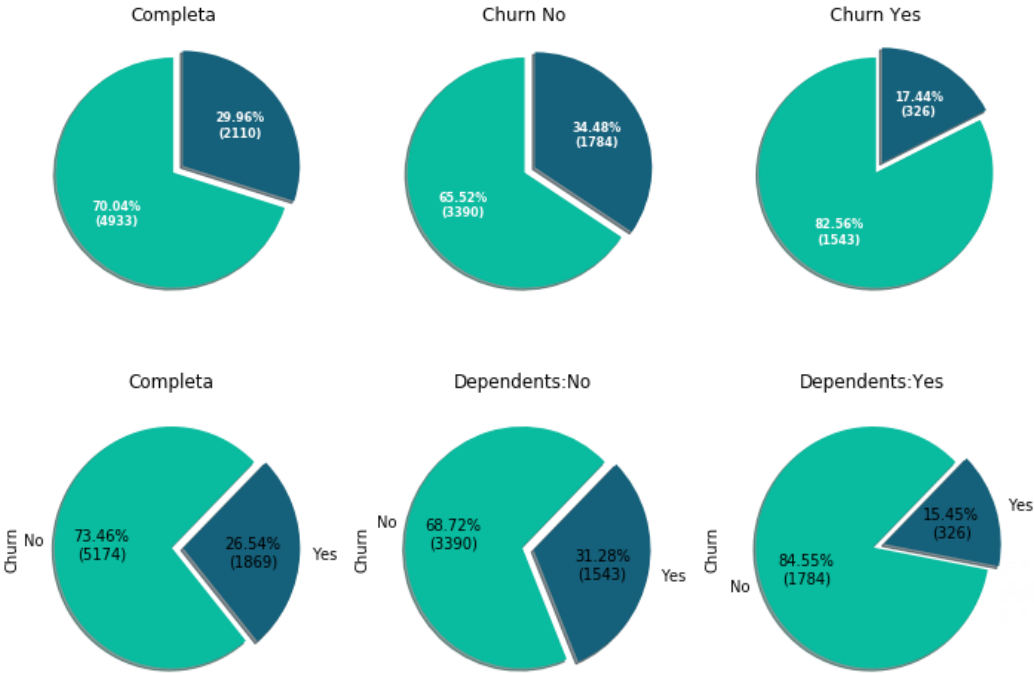


**Partner.** Sí parece variable significativa, reduciendo el churn en aquellos clientes cuyo valor de Partner es Yes

# Análisis Exploratorio. Variables categóricas

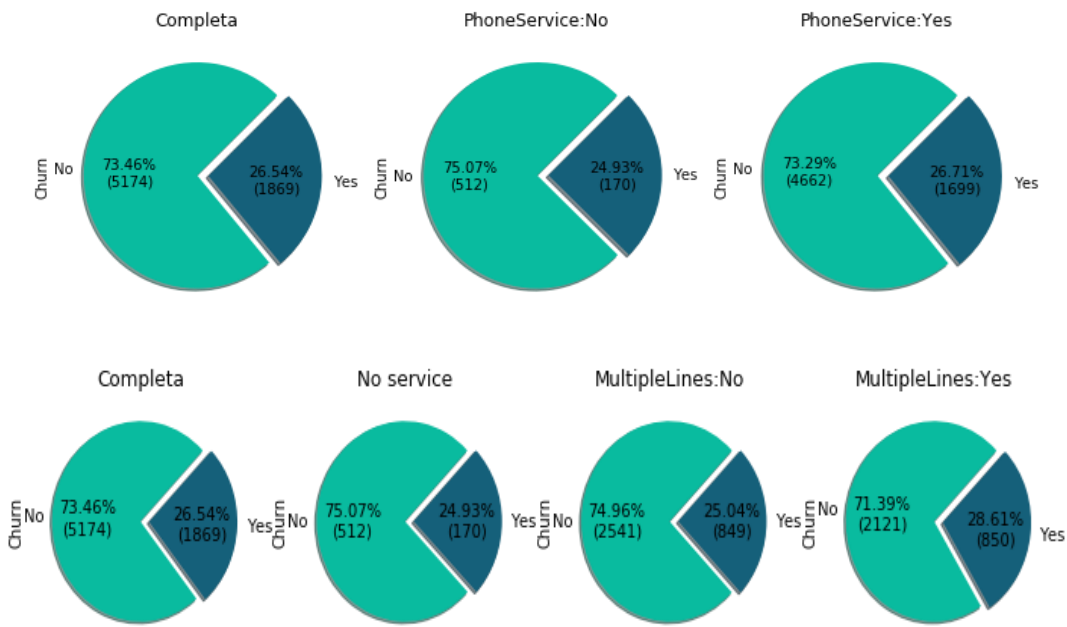


Distribucion por variable "Dependents"



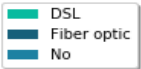
**Dependents.** Igual que Partners, sí parece reducir la tasa cuando el valor es Yes.

# Análisis Exploratorio. Variables categóricas

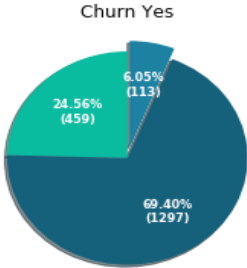
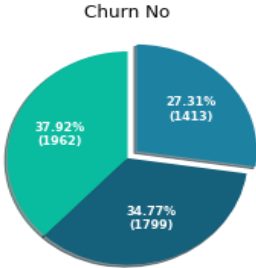
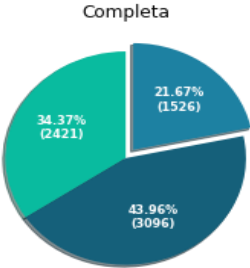


**MultipleLines.** Parece que disponer o no de múltiples líneas sí puede penalizar el churn

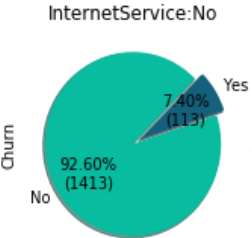
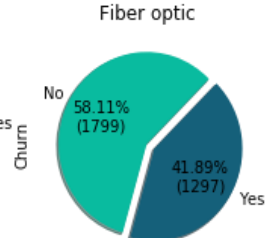
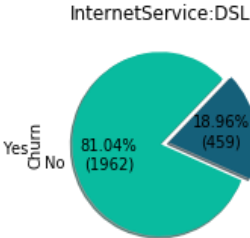
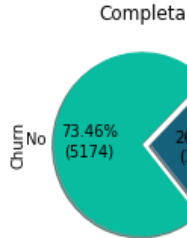
# Análisis Exploratorio. Variables categóricas



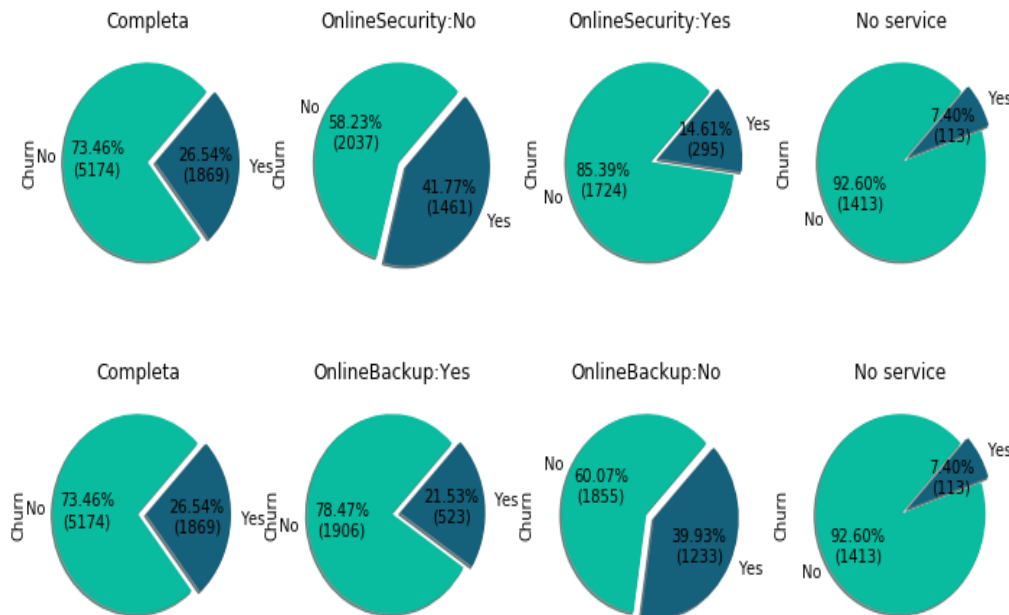
Distribucion por variable "InternetService"



**InternetService.** Parece que el servicio de fibra provoca más tasa de abandono frente al servicio de ADSL.



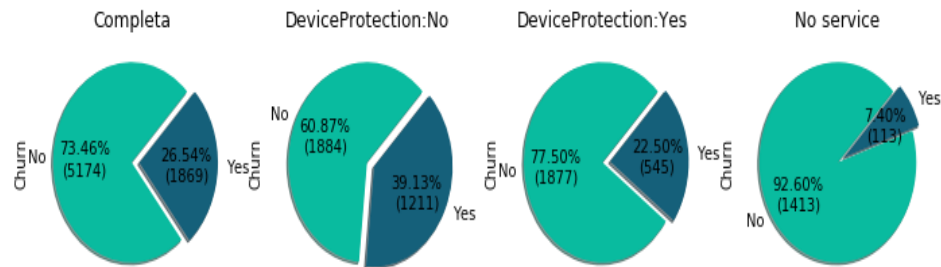
# Análisis Exploratorio. Variables categóricas



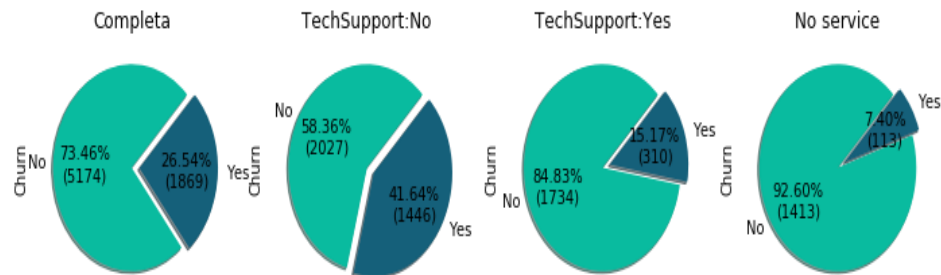
**OnlineSecurity.** Disponer de éste servicio, sí parece retener a los clientes.

**OnlineBackup.** No tanto como el anterior, disponer de éste servicio, sí parece retener a los clientes. Igual que antes, analizar el agrupar en No.

# Análisis Exploratorio. Variables categóricas

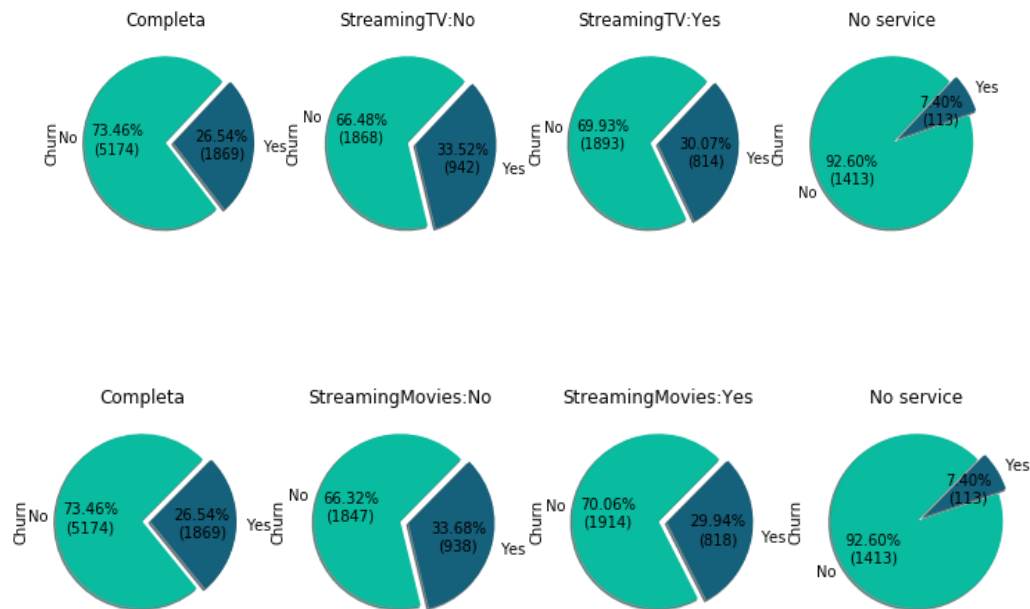


**DeviceProtection.** No igual que OnlineSecurity pero mejor que OnlineBackup, disponer de éste servicio reduce la tasa de abandono. Igual que antes, analizar el agrupar en No.



**TechSupport.** De forma análoga y con un porcentaje mejor que OnlineSecurity, disponer de éste servicio reduce la tasa de abandono. Igual que antes, analizar el agrupar en No.

# Análisis Exploratorio. Variables categóricas



**StreamingTV.** No parece que éste servicio sea uno de los mejores de la compañía.

Quizás, la razón esté en el servicio de Fibra ofrecido; por lo que puede haber una relación directa con él.

**StreamingMovies.** Muy similar a StreamingTV, no parece ser uno de los mejores servicios que proporciona la compañía.

# ICEMD

INSTITUTO ECONOMÍA  
DIGITAL

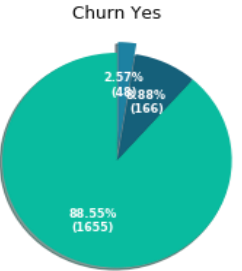
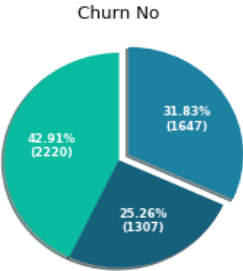
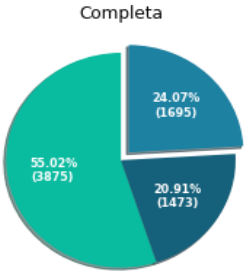


ESIC

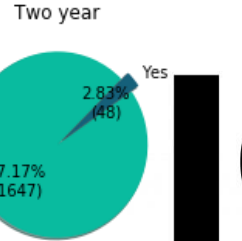
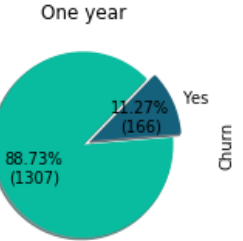
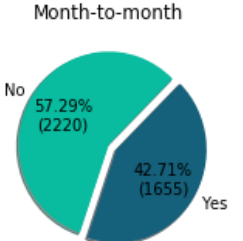
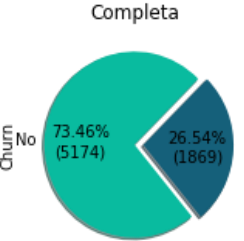


# Análisis Exploratorio. Variables categóricas

Distribucion por variable "Contract"



**Contract.** El tipo de contrato determina de una forma muy clara la tasa de abandono siendo muy pequeña en el contrato "Two years". Puede que incluso, cuanto más largo sea el tiempo del contrato, también, sea mayor la permanencia exigida al cliente.

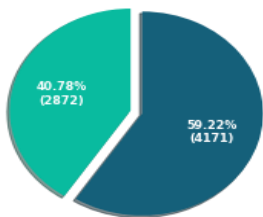


# Análisis Exploratorio. Variables categóricas

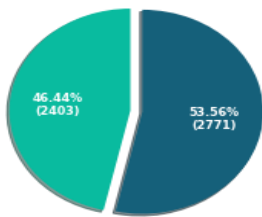


Distribucion por variable "PaperlessBilling"

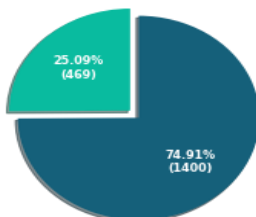
Completa



Churn No

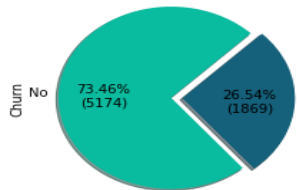


Churn Yes

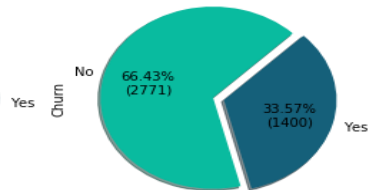


**PaperlessBilling.** ¡¡Jamás hubiera pensado que no disponer de factura electrónica podría llegar a mejorar la tasa de abandono!! ¿Puede ser que el cliente tenga más asequible la factura? En cuyo caso, ¿puede ser que la compañía tenga problemas de facturación y provoque el abandono?

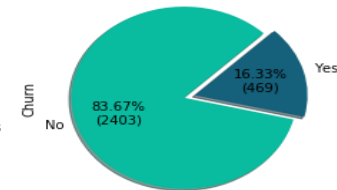
Completa



PaperlessBilling:Yes



PaperlessBilling:No



# ICEMD

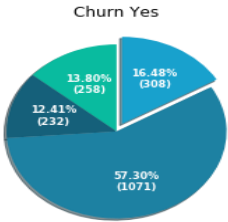
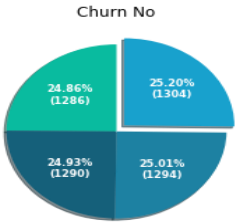
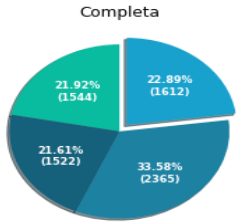
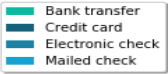
INSTITUTO ECONOMÍA  
DIGITAL



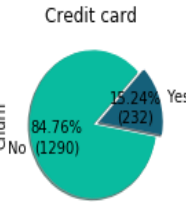
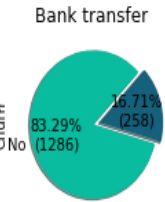
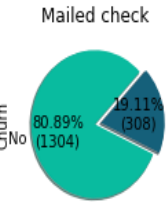
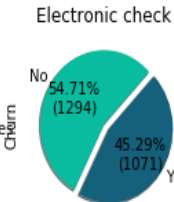
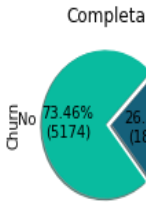
ESIC

# Análisis Exploratorio. Variables categóricas

Distribucion por variable "PaymentMethod"



**PaymentMethod.** Parece que el método de pago "Electronic check" es el que más penaliza la tasa de abandono.



# Selección de Variables

- **WOE e IV.** Es una herramienta muy poderosa porque nos proporciona unos valores para decidir qué variable debe quedar dentro o fuera (Suspicious, Strong, Medium, Weak, Useless).
- **Matriz de Correlación.** De la misma manera, se puede ver la correlación de todas las variables con la variable objetivo y decidir con cuál quedarse.
- **RFE.** Utilizando el modelo ML, se itera varias veces hasta quedarse con las variables que mejor scoring proporcionan.

Necesario discretizar las variable numéricas:

- Rango de valores.
- Percentiles + desplazamiento media.
- Algoritmos de clustering como K-Means.  
Más aconsejable **Jenks Natural Breaks**.

Para utilizar éste algoritmo, es necesario instalar un paquete pip `install jenkspy`.

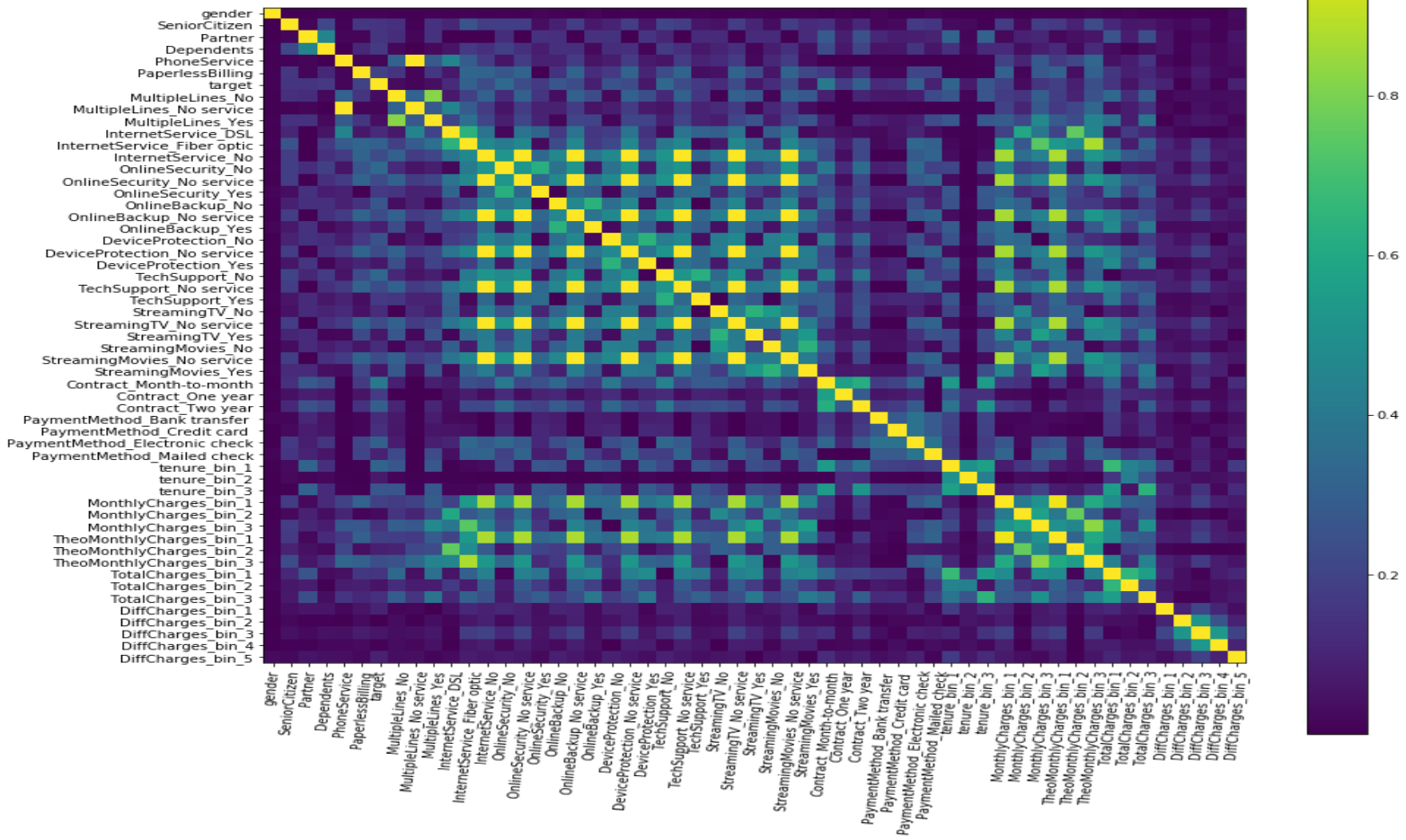
# ICEMD

INSTITUTO ECONOMÍA  
DIGITAL



ESIC

# Selección de Variables. Matriz de Correlación



# Selección de Variables.

IV. Hay variables muy predictoras como puedan ser **Contract** o **tenure**.

Así mismo, hay variables que no afectan al churn como pueda ser **gender**.

## **Matriz de Correlación.**

De forma equivalente, vemos que hay variables muy determinantes como **Contract** o **tenure** y otras mucho menos como **gender**.

## **Matriz de Correlación.**

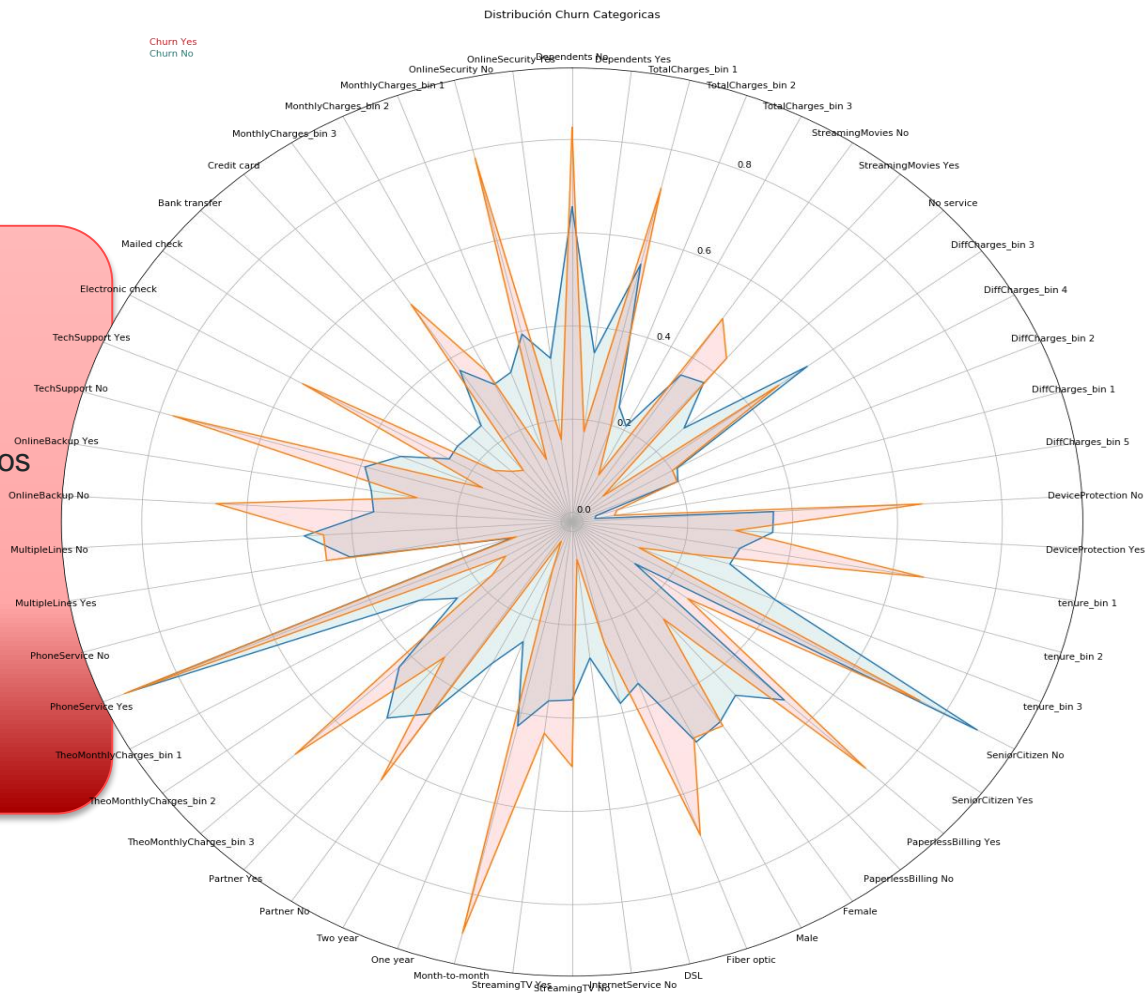
- Podemos agrupar MultipleLines en dos valores Yes/No.
- Podemos agrupar servicios de datos (Streaming, Online, etc..) en dos valores Yes/No.
- Hay ciertos servicios de datos de valor añadidos que están correlados, parece todo apuntar a que debe existir algún tipo de oferta pack de servicios aunque, como hemos visto, a nivel facturación no parece haber descuento por pack ya que cada servicio va por separado.
- La cuota mensual de 18.25 a 42.4 está asociada a servicios **No Fibra**, Voz y DLS y pocos servicios de valor añadido, tal y como vimos al calcular los precios aproximados.
- La cuota mensual de 77.8 a 118.75 está asociada a servicios de Internet **Fibra** tal y como ya habíamos visto.

# Diagrama de Afinidades

De forma general, a grandes rasgos:

- Clientes nuevos (tenure<3).
- Clientes de Fibra.
- Clientes sin servicios valor añadido contratados exceptuando Streaming.
- Clientes con contrato mensual.
- Clientes sin “Dependents” y sin “Partners”

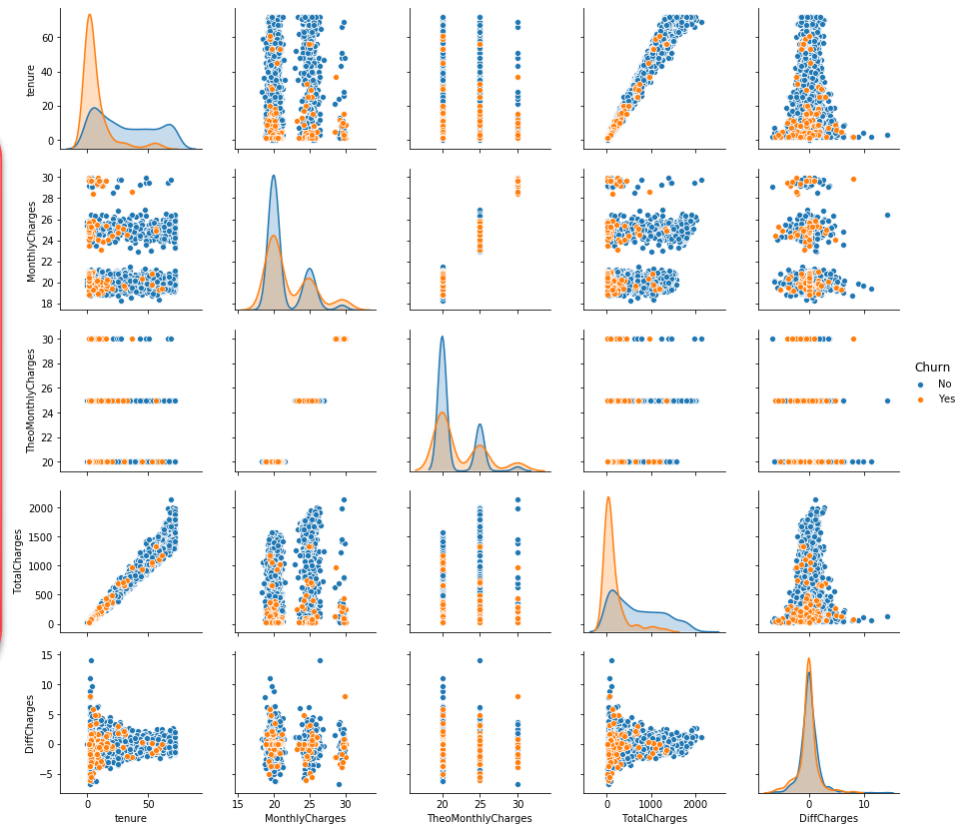
**Son clientes potenciales de Churn**



# Caso Interesante. Clientes cuota < 25

Clientes con apenas antigüedad, parecen bastante fieles.  
Clientes que contratan Fibra y, en consecuencia, algún servicio de Streaming, además, de tener una cuota superior, parece existir una tasa de abandono clara.  
Quizás, el servicio no cumple las expectativas del cliente, quizás, la competencia tiene mejor servicio / oferta.

No Yes

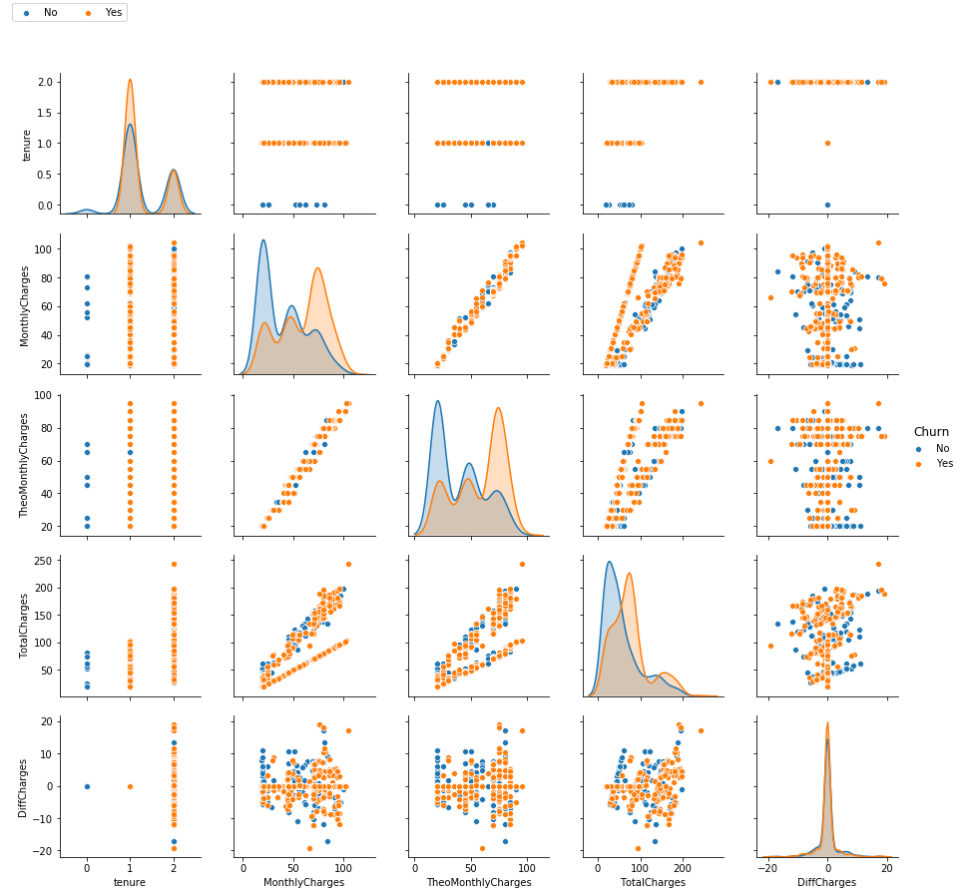


























# Caso Interesante. Clientes Nuevos

Servicios de **Fibra** junto con Contrato **month-to-month** son una "**bomba explosiva**" que favorece el **Churn**.

Interesante analizar por qué los clientes se marchan con poca vida en la compañía.



# Modelado. BigML. optiML

 2			 23			 0			 38			 2			Show metrics for: <span>All classes</span>		
All models			Max. phi coefficient			ROC AUC											
	L1 regularized (c=31.3630753756), no bias, auto-...		0.46092			0.83726											
	L2 regularized (c=20.1673831493), bias, auto-sca...		0.45923			0.83699											
	L1 regularized (c=19.8220500126), bias, auto-sca...		0.46334			0.83574											
	L1 regularized (c=1.24355793744), bias, auto-sca...		0.48699			0.83533			 <input type="checkbox"/> ▼								
	boosted trees, 214-node, 20-iteration, determinist...		0.47146			0.83532			 <input type="checkbox"/> ▼								
	L1 regularized (c=22.9594629034), bias, auto-sca...		0.48641			0.83382			 <input type="checkbox"/> ▼								
	L1 regularized (c=17.9026849344), no bias, auto-...		0.48502			0.83339			 <input type="checkbox"/> ▼								
	bootstrap decision forest, 242-node, 23-model, d...		0.47624			0.83236			 <input type="checkbox"/> ▼								
	L1 regularized (c=31.7838507965), no bias, auto-...		0.48208			0.83097			 <input type="checkbox"/> ▼								
	L2 regularized (c=1), bias, auto-scaled, missing va...		0.45828			0.82418			 <input type="checkbox"/> ▼								

65 modelos entrenados:

- **Regresión Logística** encabeza los cuatro primeros (con un ROC AUC de 0.83726).
- Un modelo **Random Forest** (ROC AUC de 0.83532) ocupa el 5º lugar.
- Deep Learning queda muy a la cola
- Un **Arbol de Decisión** queda al final (ROC AUC 0.72135).

# Modelado. BigML. optiML

Usaremos algunos de los modelos vistos en clase.

**SVN** lo dejaremos fuera por razones de tiempo computacional.

Para optimización, usaremos las métricas **f1\_score** y **roc\_auc**.

Consideramos **Churn = 1** como positivo.

Consideraremos el DataSet codificado en LabelEncoding y OneHotEncoding.

Utilizaremos varios métodos de OverSampling y UnderSampling

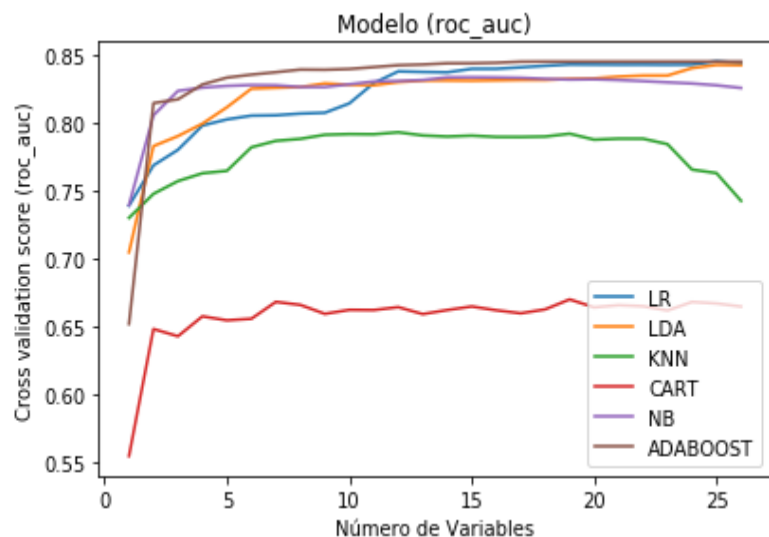
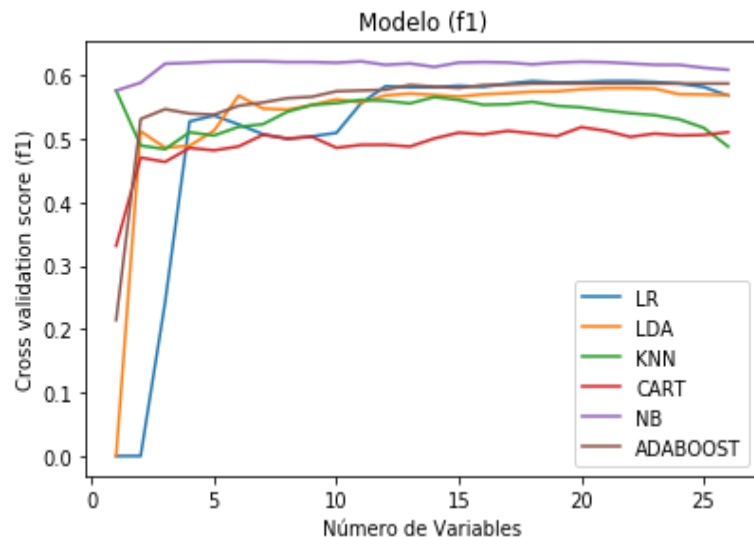
Probaremos técnicas de mejora de Hyper Parámetros.

```
scorings = ['f1', 'roc_auc']
```

```
samplers = [  
    ('RandomOverSampler', RandomOverSampler()),  
    ('SMOTE', SMOTE()),  
    ('ADASYN', ADASYN()),  
    ('RandomUnderSampler', RandomUnderSampler()),  
    ('NearMiss', NearMiss()),  
    ('SMOTEENN', SMOTEENN())  
]
```

```
models = {  
    'LR': LogisticRegression(),  
    'LDA': LinearDiscriminantAnalysis(),  
    'KNN': KNeighborsClassifier(),  
    'CART': DecisionTreeClassifier(),  
    'NB': GaussianNB(),  
    'ADABOOST': AdaBoostClassifier()  
}
```

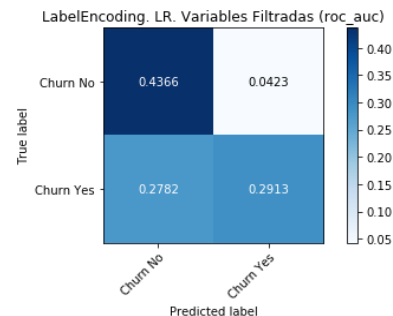
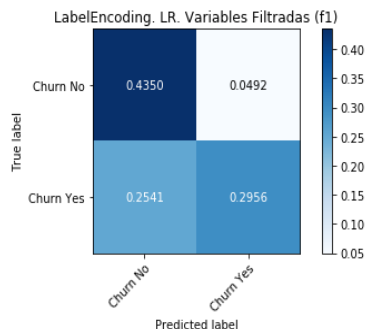
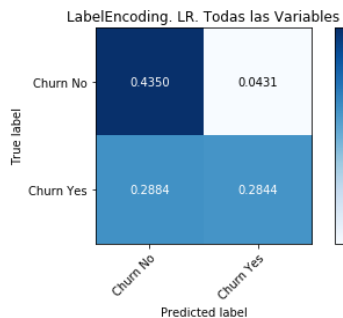
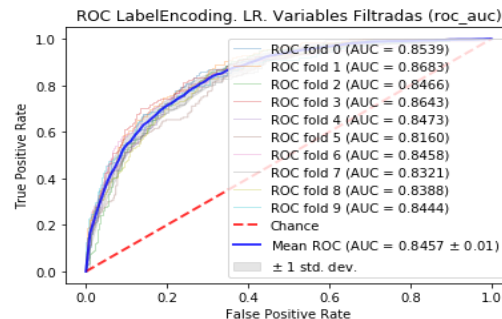
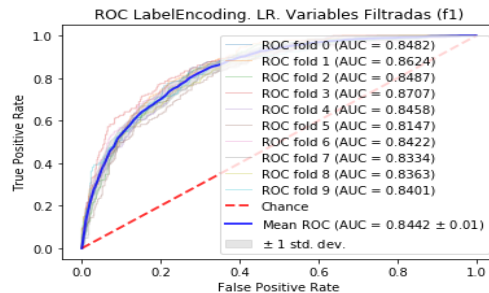
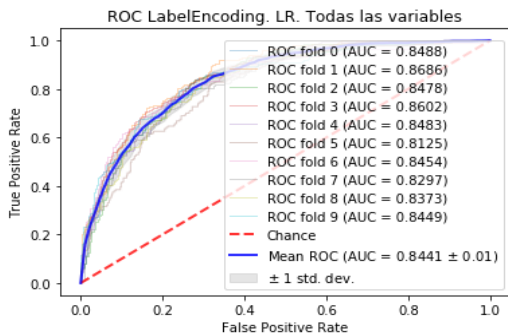
# LabelEncoding



LR: Variables Eliminadas

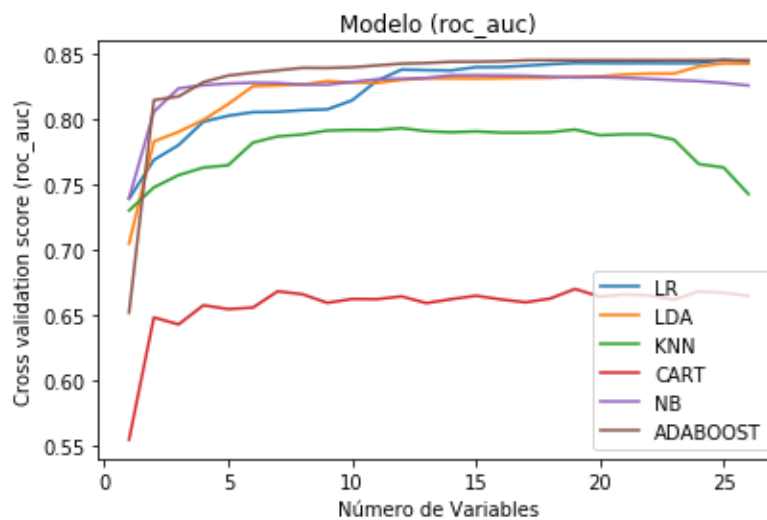
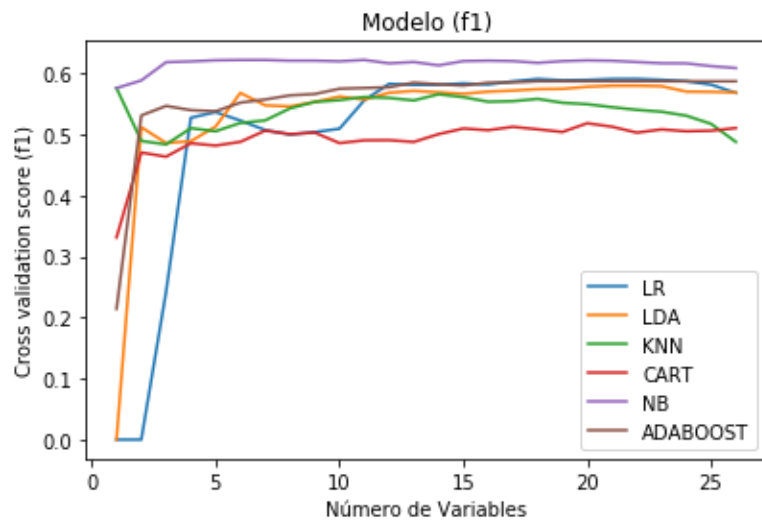
f1 => ['DiffCharges', 'MonthlyCharges', 'TheoMonthlyCharges\_bin', 'TotalCharges']  
roc\_auc => ['DiffCharges']

# Regresión Logística. LabelEncoding



Los ratios mejoran si se utilizan sólo las variables más importantes.  
Las variables seleccionadas por **RFE** difieren mucho de **IV**.  
Sin embargo, la mayoría de los modelos tienen una relación entre los TP y FN muy similar, casi el **50%**

# One Hot Encoding (get\_dummies)

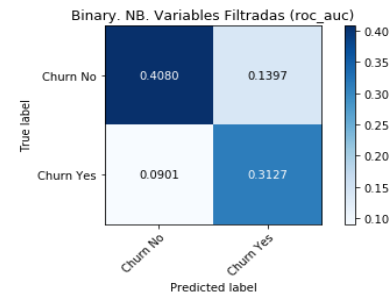
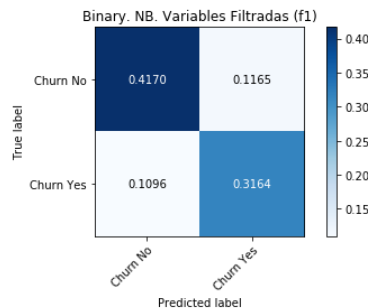
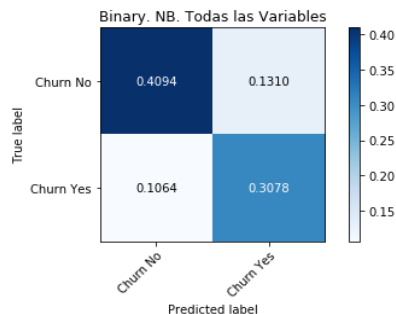
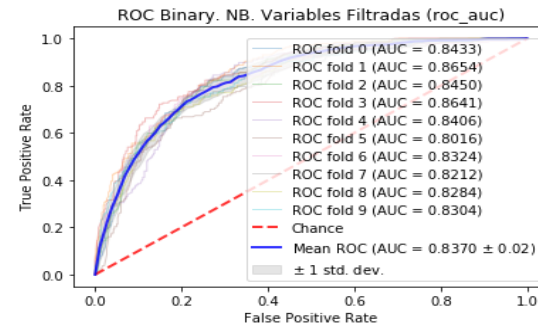
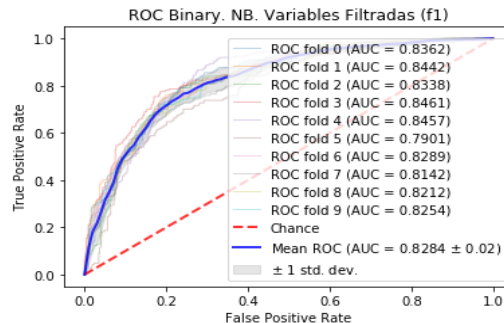
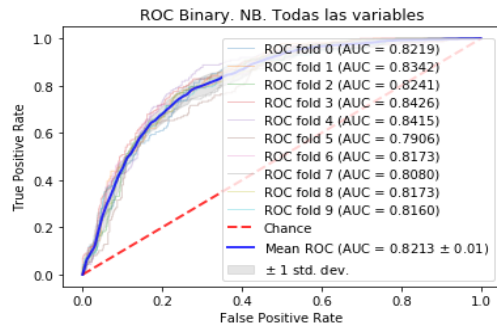


ADABOOST: N° Optimo de Variables

f1 => 30 de 40

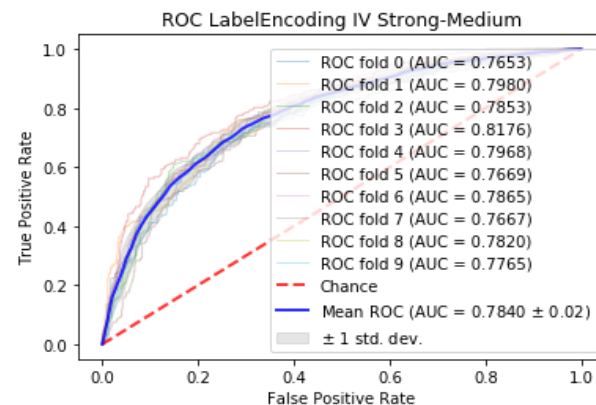
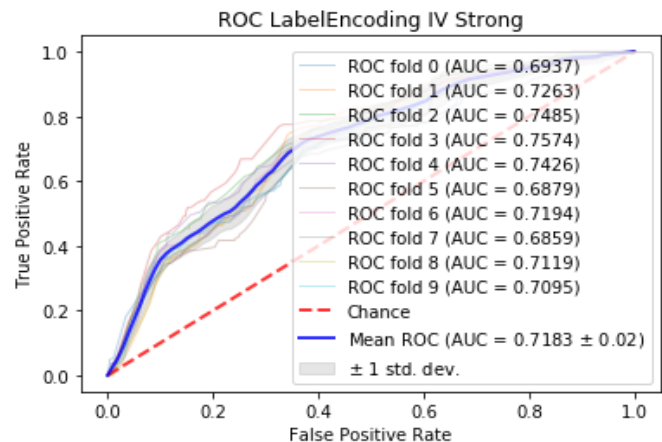
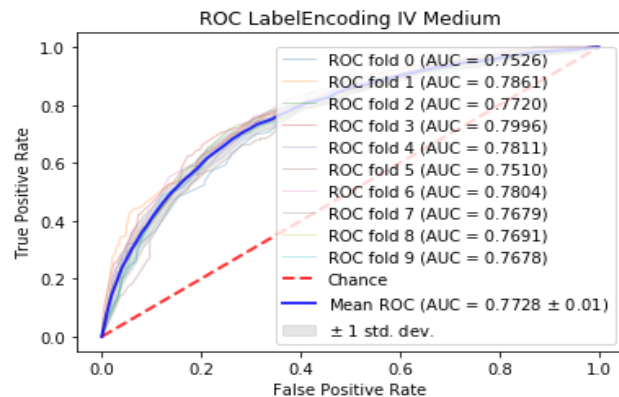
roc\_auc => 21 de 40

# GaussianNB. One Hot Encoding



De todos los modelos, NB es el que mejor funciona de cara a predecir TP.  
Seguramente, habrá un **modelo mixto**.

# Regresión Logística. WOE Features

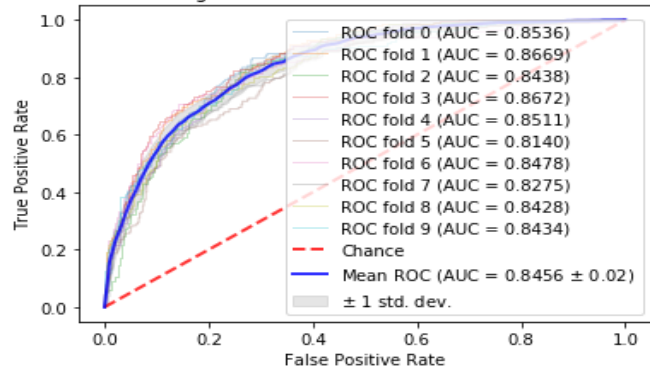


Las variables seleccionadas por **RFE** dan mucho mejor resultado que las proporcionadas por **IV**.

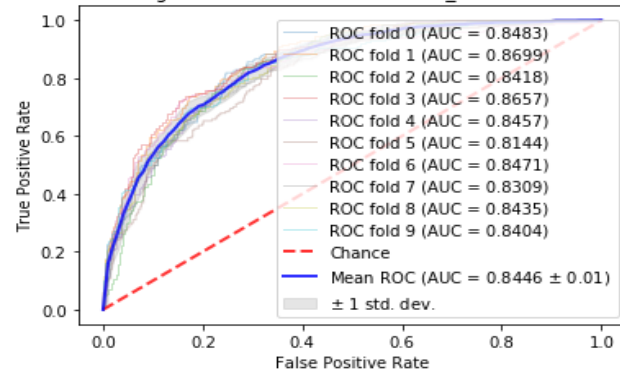


# Regresión Logística. OverSampling vs UnderSampling

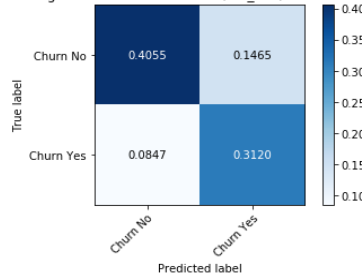
ROC LabelEncoding. LR. Variables Filtradas (f1) RandomOverSampler



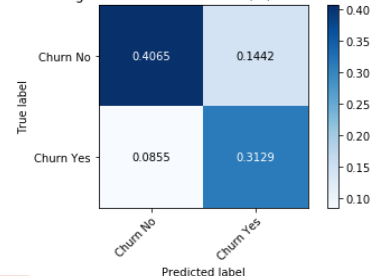
ROC LabelEncoding. LR. Variables Filtradas (roc\_auc) RandomUnderSampler



LabelEncoding. LR. Variables Filtradas (roc\_auc) RandomUnderSampler



LabelEncoding. LR. Variables Filtradas (f1) RandomOverSampler

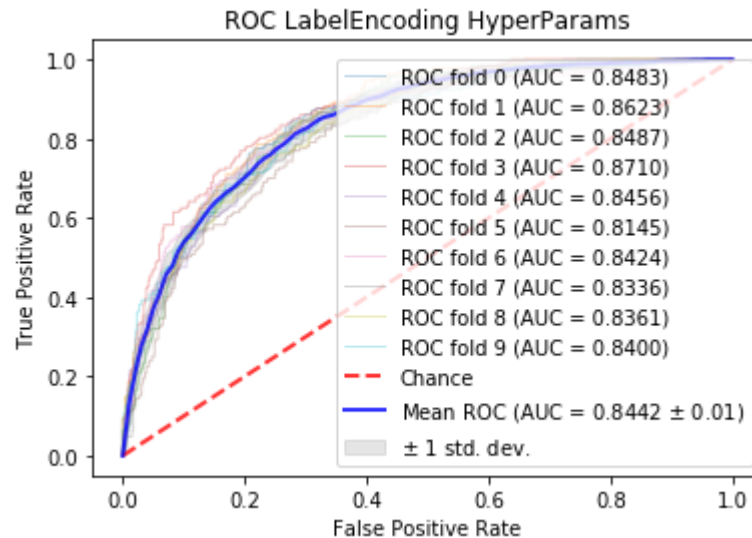
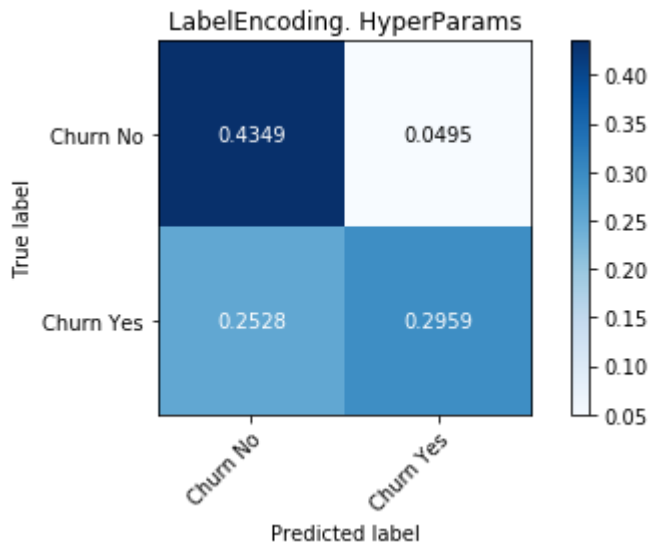


Utilizando éstas técnicas mejoran los coeficientes.  
También, mejora la relación entre los TP y los FN.  
Sin embargo, cada modelo mejora con una técnica específica

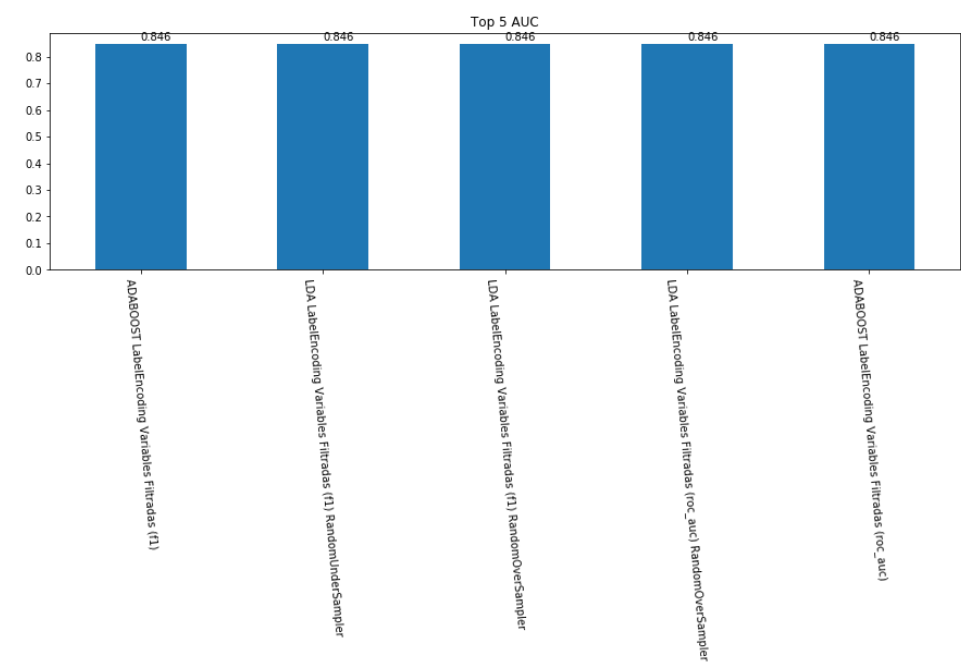
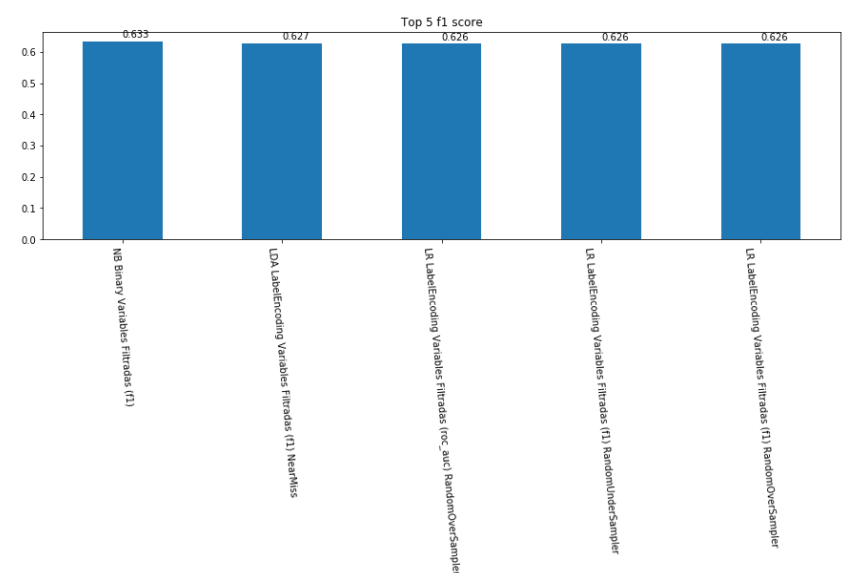
# Regresión Logística. Hyper Parámetros

Arma muy potente pero muy compleja:

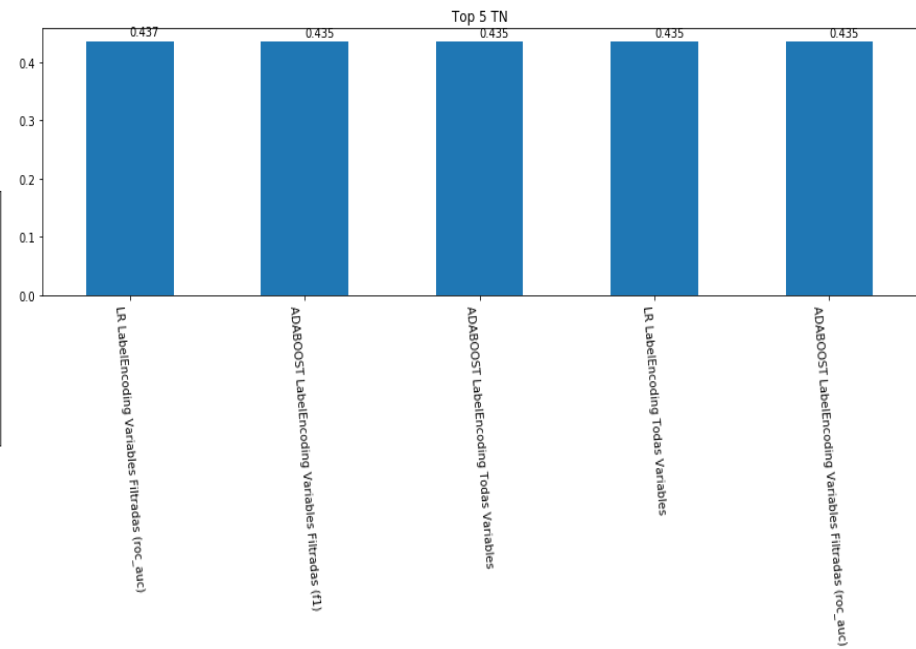
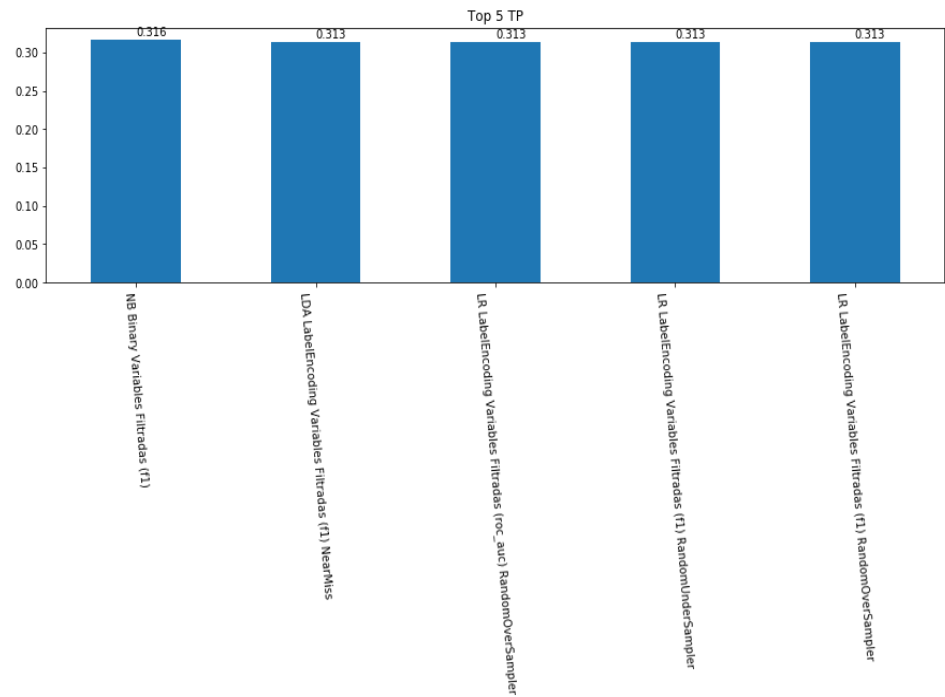
- Relación entre parámetros.
- Cross Validation.
- Más de una métrica seleccionada.
- Más de un DataSet codificado.



# Comparativa Modelos



# Comparativa Modelos



# Selección Modelo

La principal problemática es que la relación de los TP y los FN es casi la misma (50%).

En entorno productivo, introduciría lógica añadida **if else** con una secuencia similar a:

1. Usar el algoritmo que mejor TN/FP proporcione.
2. Si la probabilidad de predicción TN de ese cliente es buena => Etiquetarlo como **Churn=No**.
3. Si la probabilidad de predicción TN de ese cliente es mala => Utilizar el algoritmo que mejor TP/FN proporcione.
4. Si la probabilidad de predicción TP de ese cliente es buena => Etiquetarlo como **Churn=Yes**.
5. Si la probabilidad de predicción TP de ese cliente es mala => (ante la duda) Etiquetarlo como **Churn=Yes**.

- \* Mejor **f1\_score** => NB. Binary. Variables Filtradas (f1)
- \* Mejor **roc\_auc** => ADABOOST. LabelEncoding. Variables Filtradas (f1) ¿?
- \* Mejor **TP** => NB. Binary. Variables Filtradas (f1)
- \* Mejor **TN** => LR. LabelEncoding. Variables Filtradas (roc\_auc)

# Conclusiones

RFE funciona mejor que IV/WOE (**gender**, **contract..**)

Análisis exploratorio para “**olisquear**” más que necesario.

**Label Encoding** vs **One Hot Encoding** depende del modelo

**Cross Validation** cambia la foto de forma radical

La creación de **variables extras** es todo un arte, y en este caso, hemos tenido suerte.

Hacer **tuning** de los hiper parámetros hace que pueda mejorar el modelo, sin embargo es tarea harto complicada.

En ambos proyectos (Grupal / Individual) disponer de un **framework** de desarrollo con funciones parametrizables son la clave del éxito en un entorno de mucha **prueba** y **error**.

¿La secuencia de ejecución es la óptima?

1. Selección de **métricas** para optimizar en el **Cross Validation**.
2. Selección de **variables** basadas en esas métricas.
3. Tuning del **DataSet** para intentar mejorar esas métricas (OverSampling, UnderSampling).
4. Tuning del **Modelo** para intentar mejorar esas métricas (RandomSearch)

# Gracias

---

[https://github.com/jazzphoenix/icemd\\_bigdata](https://github.com/jazzphoenix/icemd_bigdata)

